Performance of supertree methods for estimating species trees

A thesis submitted in partial fulfilment of the requirements for the

Degree

of Master of Science in Statistics

in the University of Canterbury

By Yuancheng Wang

University of Canterbury,

Christchurch, New Zealand.

July 2010

Table of Contents

Acknowledgements

I would like to thank everyone along the journey of this research. In particular, I would take this opportunity to thank my senior supervisor, James H. Degnan and co-supervisor, Carl Scarrott for their patient guidance and useful comments. I enjoyed collaborating with them.

Then I would like to thank all the member of my department, Mathematics and Statistics in the University of Canterbury, who helped me directly or directly. Especially, Paul Brouwers and Steve Gourdie in the department for their IT support so that the simulation study is carried out successfully and smoothly.

Finally, I wish to thank my parents for their kind and endless support both spiritually and financially throughout my entire study. I owe much gratitude to them, without whose assistance this would never happen.

Abstract

Phylogenetics is the research of ancestor-descendant relationships among different groups of organisms, for example, species or populations of interest. The datasets involved are usually sequence alignments of various subsets of taxa for various genes. A major task of phylogenetics is often to combine estimated gene trees from many loci sampled from the genes into an overall estimate species tree topology. Eventually, one can construct the tree of life that depicts the ancestor-descendant relationships for all known species around the world. If there is missing data or incomplete sampling in the datasets, then supertree methods can be used to assemble gene trees with different subsets of taxa into an estimated overall species tree topology.

In this study, we assume that gene tree discordance is solely due to incomplete lineage sorting under the multispecies coalescent model (Degnan and Rosenberg, 2009). If there is missing data or incomplete sampling in the datasets, then supertree methods can be used to assemble gene trees with different subsets of taxa into an estimated species tree topology. In addition, we examine the performance of the most commonly used supertree method (Wilkinson et al., 2009), namely matrix representation with parsimony (MRP), to explore its statistical properties in this setting. In particular, we show that MRP is not statistically consistent. That is, an estimated species tree topology other than the true species tree topology is more likely to be returned by MRP as the number of gene trees increases. For some situations, using longer branch

lengths, randomly deleting taxa or even introducing mutation can improve the performance of MRP so that the matching species tree topology is recovered more often.

In conclusion, MRP is a supertree method that is able to handle large amounts of conflict in the input gene trees. However, MRP is not statistically consistent, when using gene trees arise from the multispecies coalescent model to estimate species trees.

1 Introduction

1.1 Research objective

In phylogenetics, evolutionary trees describe the heritage of species or populations, Estimated species tree can be constructed from gene trees which are sampled at different loci. One issue is that not all the genes are always sampled for each taxon, (i.e. missing data or incomplete sampling). If there is little missing data, a feasible approach to this problem is consensus methods where one only keeps genes sampled for all taxa. However, for real-world datasets, this may exclude useful information.

Supertree methods on the other hand are suitable for datasets with incomplete sampling. The supertree method called matrix representation with parsimony (MRP) is studied in this thesis. MRP was originally proposed independently by Baum (1992) and Ragan (1992). Since then, there has been debate about the properties of MRP. Its main advantages are: apparently accurate and well resolved supertree; parsimony is a familiar and well-understand optimization (Bininda-Emonds, 2004). However, its main disadvantages are: non-independence of input trees and weighting nodes in source phylogenies (Gatesy and Springer, 2004).

As the numbers of input trees or shared taxa increases, it is common to have conflict among the input trees. Essentially, MRP weights the evidence in different input trees

to handle the conflict, without having any input trees overwhelming the rest. With the aid of modern computer power, we can explore the properties of MRP from new perspectives.

In this project, we will investigate the consistency of MRP for inferring species trees from gene trees so that MRP is statistically consistent if and only if the probability of returning the matching species tree goes to 1 as the number of gene trees goes to infinity. We take 2 approaches to achieve this goal, namely a simulation and an analytical study. For simplicity, we only study selected bifurcating true species trees in this study. That is, each node of the tree can only have exactly 2 descendants.

1.2 Outline

Chapter 2 gives some background information about the phylogenetics field, the ideas of species tree and gene tree, some methods available in the literature and an example to demonstrate how MRP works.

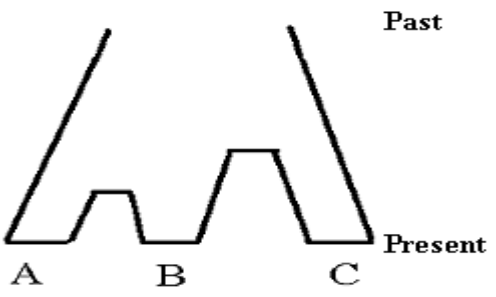
Chapter 3 describes the simulation procedure and the idea of an outgroup.
It also discusses the mutation model and pruning schemes used in the simulation and how to assess the performance of MRP. This chapter ends by several applications of 4-taxon, 5-taxon and 20-taxon species trees and results to illustrate the performance of MRP under difference simulation settings.

Chapter 4 explains the idea of the expected parsimony score and its implementation. It provides some applications to 4-taxon and 5-taxon in difference circumstances that demonstrate the conditions for which MRP is consistent and also some guidelines of the pruning schemes. Finally, the summary is given in Chapter 5 to highlight the key results.

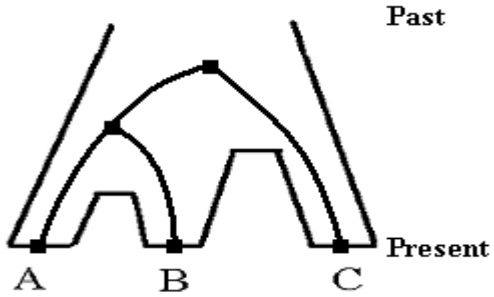
2 Background

2.1 Introduction of phylogenetics

Phylogenetics is a tool to gain an insight into the ancestor-descendant relationship of species through a variety of methods. An approach is to infer a species tree using a set of input gene trees sampled from an underlying true species tree, which is unknown in general. The species tree is used to describe the ancestor-descendant relationships of a set of populations, whereas the ancestor-descendant relationships for the same gene sampled from several individuals in different populations is denoted by the gene tree. The tips of both species trees and gene trees are called taxa. The set of relationships between all taxa is represented by the tree topology.



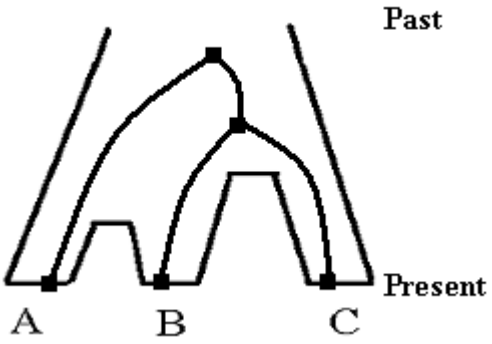
An example species tree with 3 taxa A, B and C.
Height and width of the species tree represents the generations and the population size respectively.
Notice that populations A and B have a closer relationship to each other than to C.



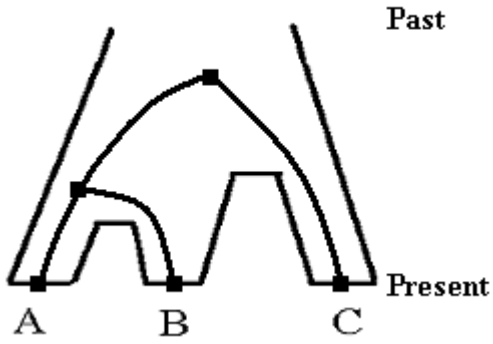
An example gene tree that is sampled from the species tree at the left.
Lineages A and B join first at a node going backwards in time before joining with C, which indicates that lineages A and B are more closely related to each other than to lineage C.
Nodes of gene tree represent coalescence event where 2 genes are copied from the same ancestral gene.

Figure 1. Species tree and gene tree with the same topology ((A,B),C).

Gene tree discordance occurs where the topology of the gene tree is different from the topology of the true species tree. The source of gene tree discordance can be incomplete lineage sorting; horizontal gene transfer; gene duplication and loss; hybridization and recombination. In this research, we assume incomplete lineage sorting as the sole cause of gene tree discordance. In incomplete lineage sorting, 2 or more lineages fail to coalesce within the most closely related populations first. This makes it possible for at least 1 of the lineages to coalesce with a lineage that comes from a less related population.



An example of incomplete lineage sorting.
The lineage from population B coalesces with a less related lineage from population C first going backwards in time instead of the lineage from population A, even though population A and B are the most closely related in this example.



An example where incomplete lineage sorting does not occur.
The lineage from population B coalesces with lineage from the most closely related population A first going backwards in time.

Figure 2. Incomplete lineage sorting.

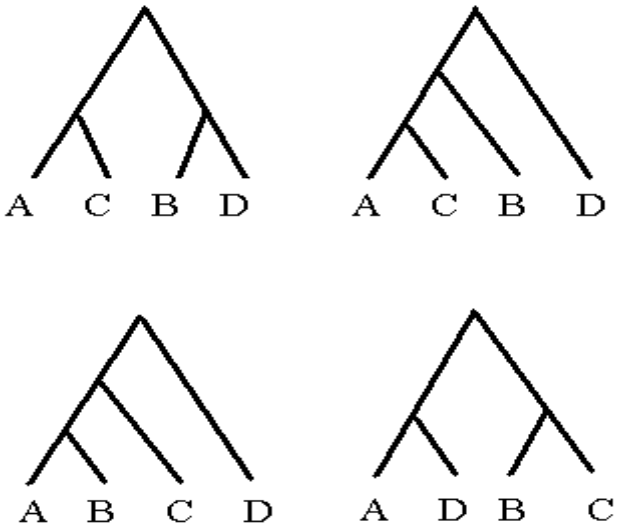
Consensus and supertree are 2 phylogenetics methods that can be used to infer species trees. Both methods summarise the relationships within a set of input trees. The main difference is that supertree methods only require the input trees to have overlapping

taxa sets, whereas consensus methods need every input tree to have the same taxa set. Hence, supertree methods are a generalisation of consensus methods by allowing different subsets of taxa for different gene trees. In practice, it is not always feasible to sample every single taxon for all genes of interest. Therefore, researchers in the phylogenetics field often prefer supertree methods. Consensus methods only work when every taxon is sampled for every gene (i.e. no missing data).

2.2 Consensus methods

Majority-rule consensus and greedy consensus are examples of simple consensus methods to construct a tree from input trees with the same set of taxa. Both algorithms use the probabilities of clades to infer trees. A clade is a subset of at least 2 taxa that is monophyletic, meaning most recent common ancestor of the taxa is not ancestral to any other taxa on the tree. The majority-rule consensus method only uses those clades that occur more than 50% of the time in the input trees (Felsenstein, 2004). However, the greedy consensus method assembles a tree by gradually adding the most frequent clade which is compatible with clades that already accepted in the tree (randomly breaking the ties) (Bryant, 2003). The greedy consensus method is also known as the majority-rule extended consensus method (Degnan et al., 2009).

Figures 4 and 5 are used to illustrate how majority-rule consensus and greedy consensus work with the same input trees, in Figure 3.



Existed Clade	Probability
{AB}	1/4
{AC}	1/2
{AD}	1/4
{ABC}	1/2
{ABCD}	1
{BC}	1/4
{BD}	1/4

Figure 3. Input trees and clades probabilities.
Clade {AC} occurs twice from the input trees (top row),
its corresponding clade probability is 2/4 = 1/2 and so forth.

{ABCD} is the only clade with probability bigger than 1/2 in the list.

Therefore, an unresolved tree is returned by the majority-rule consensus, also known as the star tree.

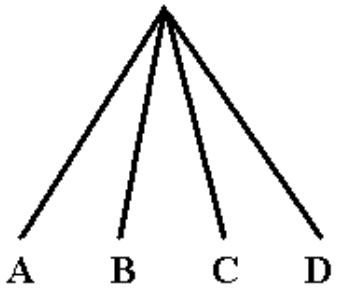
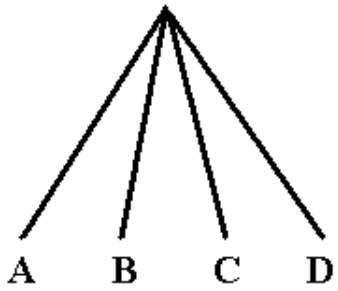


Figure 4. Majority-rule consensus approach.

{ABCD} is the most frequent and compatible clade in the list.

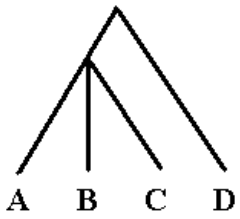
An unresolved tree is returned by the greedy consensus first.



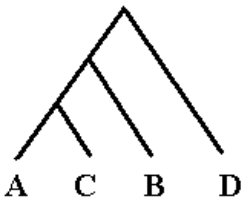
Because both {AC} and {ABC} are the most frequent and compatible clades (a tie) at this stage, 1 of them is picked randomly.

Case 1
Greedy consensus picks clade {ABC} first:

{ABC} is the most frequent clade from the list and compatible with {ABCD} clade.

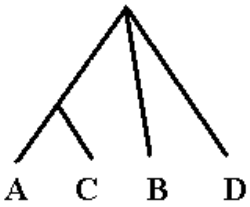


Then, {AC} is the most frequent clade in the list and compatible with {ABC} clade.

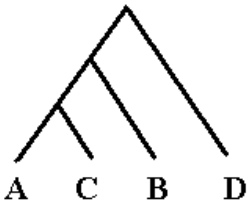


Case 2
Greedy consensus picks clade {AC} first:

{AC} is the most frequent clade from the list and compatible with {ABCD} clade.



Then, {ABC} is the most frequent clade in the list and compatible with {AC} clade.



The same tree is returned in both cases by greedy consensus in this example. However, this is not true in general.

Figure 5. Greedy consensus approach.

2.3 Supertree methods

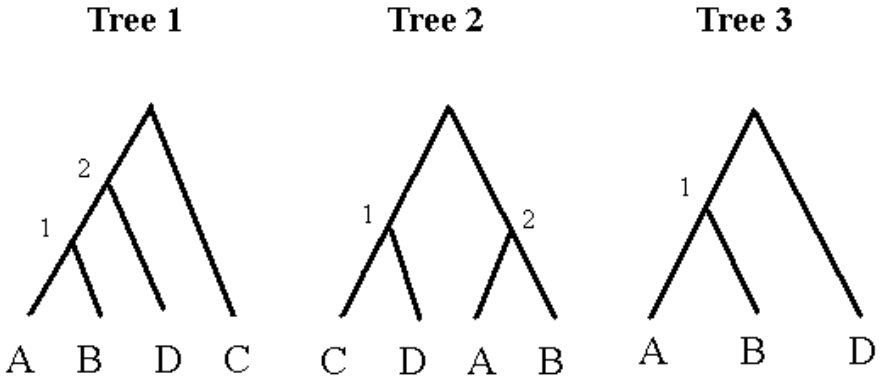
The most popular supertree method is matrix representation with parsimony (MRP)

(Wilkinson et al., 2009). The idea is basically to represent each input tree into a matrix

form. Each column of the matrix records the taxa descended from a particular internal node of the input tree, and each row represents a distinct taxon (Sanderson et al., 1998). Let M denote the matrix representation block of a tree where M_{ij} is the entry at row i and column j. Under this matrix representation, a question mark "?" is used to denote those taxa that do not occur in the current input tree (i.e. missing data); $M_{ij} = 1$ if taxon i is descended from node j of the current input tree and $M_{ij} = 0$ otherwise. The original tree structure and its matrix representation form have a one-to-one correspondence (Bininda-Emonds, 2004). A row of 0's is added at the end, which is same as adding an outgroup (see Section 3.5).

For k input trees, there are $M_1, M_2, ..., M_k$ such matrix blocks. Let $M = [M_1 \mid M_2 \mid ... \mid M_k]$ denote the overall matrix block such that each row represents a unique taxon and each column records the descendant of a node from an input tree. Let $S = \bigcup_{i=1}^{k} s_i$ where s_i is the taxa set of the input tree i.

If the full taxon set has size $n = |S|$ taxa, for the ith input tree with $n_i \leq n$ taxa, such input tree contributes $n_i - 2$ columns to the overall matrix representation M because there are exactly $n_i - 2$ internal nodes (not including the roots) as demonstrated in Figure 6 below.



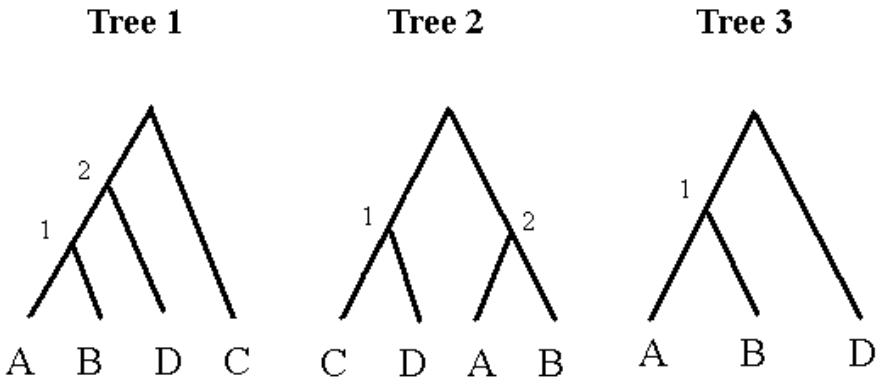
nodes / taxon	1	2
A	1	1
B	1	1
D	0	1
C	0	0
Outgroup	0	0

nodes / taxon	1	2
A	0	1
B	0	1
D	1	0
C	1	0
Outgroup	0	0

nodes / taxon	1
A	1
B	1
D	0
C	?
Outgroup	0

nodes / taxon	Tree one		Tree two		Tree
	1	2	1	2	three
A	1	1	0	1	1
B	1	1	0	1	1
D	0	1	1	0	0
C	0	0	1	0	?
Outgroup	0	0	0	0	0

Figure 6. Coding input gene trees.



Matrix representation of the three input trees:

A	1	1	0	1	1
B	1	1	0	1	1
D	0	1	1	0	0
C	0	0	1	0	?
Outgroup	0	0	0	0	0

Figure 7. Input trees and the corresponding matrix representation form.

After coding all the input trees into the matrix form, a tree (or set of tied trees) is returned by MRP if and only if its parsimony score total is the lowest among all possible candidate trees that contain the full taxa set *S*. For the example in Figure 7, tree 1 has taxa A, B, C and D; tree 2 has A, B, C and D; tree 3 has A, B and D. Hence, the union taxon set or the full taxa set *S* is A, B, C and D, for which there are 15 possible topologies (Felsenstein, 2004), see Figure 8. The parsimony score indicates that the output tree has the fewest evolutionary changes or in other words, it is the most parsimonious. Equivalently, there are as few evolutionary events as possible needed to yield the input trees from the output tree (Felsenstein, 2004). Figure 9 is an example of how to calculate the parsimony score.

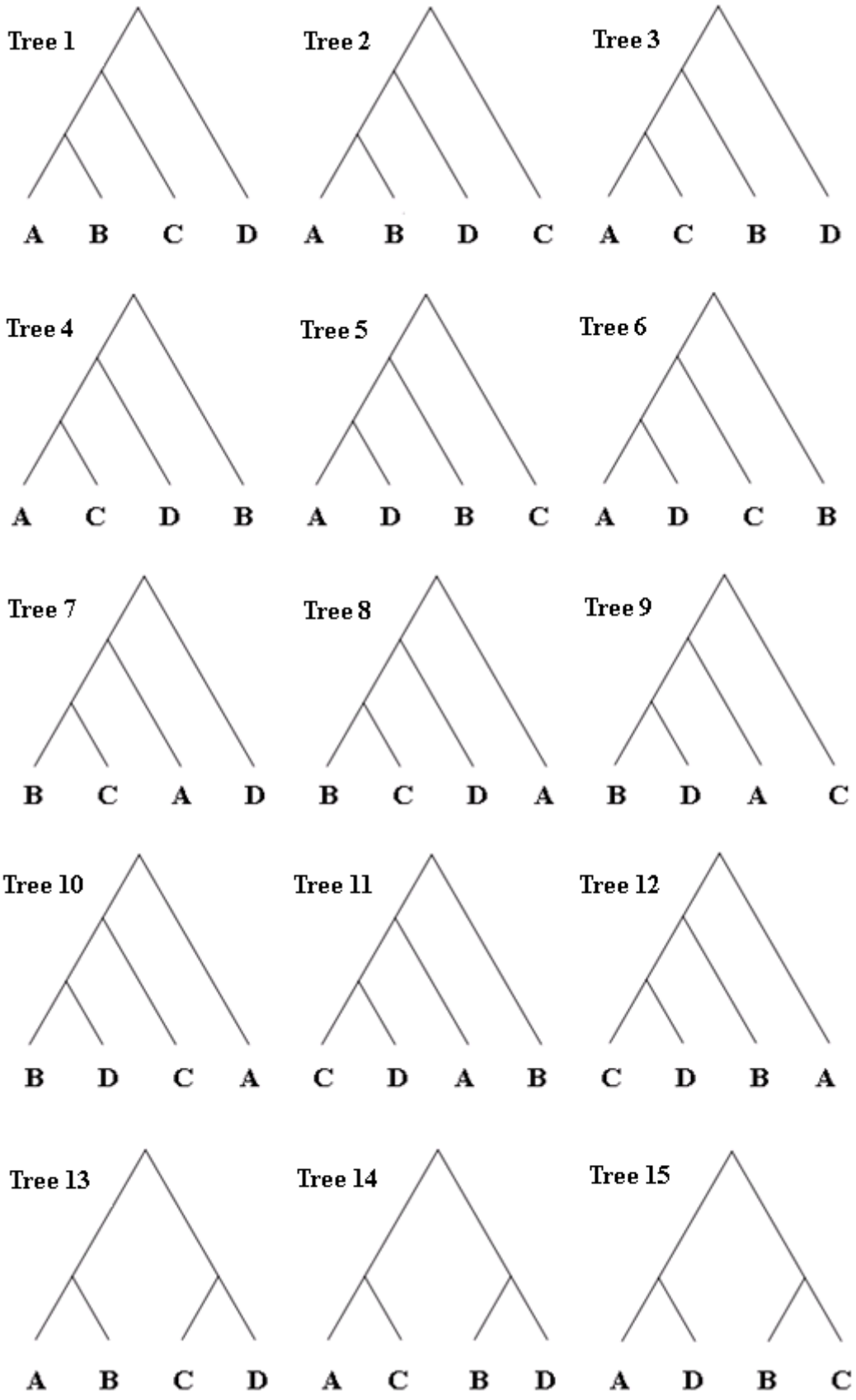
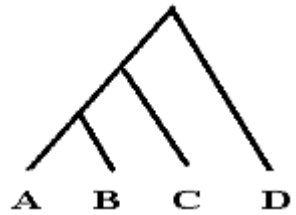


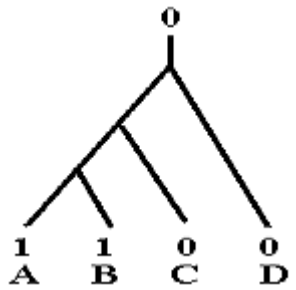
Figure 8. All 15 possible topologies with taxa set: A, B, C and D.

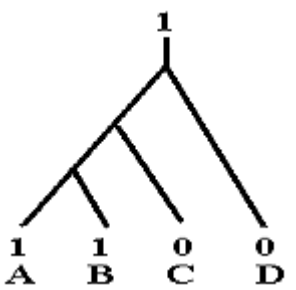


A candidate tree.

A	1
B	1
D	0
C	0
Outgroup	0

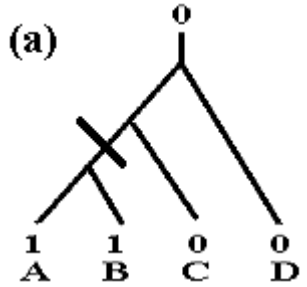
The first column of Figure 7.



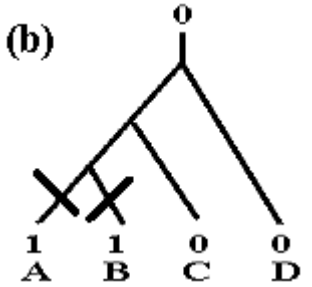


Using a 0 or 1 to denote the root of the candidate tree at the top, so that one needs to find the parsimonious (fewest) changes to match exactly the 1's and 0's at tips of the candidate tree.

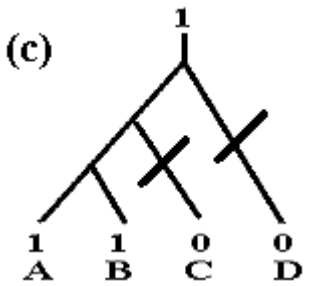
Let calculate the parsimony score for the first column of Figure 7 against the candidate tree. Notice that not all the possible cases are covered in (a) to (c) below.



Need 1 change of 0 → 1, at somewhere along the marked edge in (a).
The parsimony score is 1.



Need 2 changes of 0 → 1, at somewhere along the marked edge in (b).
The parsimony score is 2.



Need 2 changes of 1 → 0, at somewhere along the marked edge in (c).
The parsimony score is 2.

Therefore, the parsimonious change is 1 as illustrated in (a), so that the corresponding parsimony score for the first column of Figure 7 against the given candidate tree is 1.

Figure 9. Parsimony score.

For the example given in Figure 7, the MRP output tree is found by calculating the parsimony scores for all the input trees against all the 15 candidate trees (Figure 8) and picking the tree with the lowest parsimony score total as the output tree.

Input tree / Candidate tree	Tree 1		Tree 2		Tree 3	Total score
1	1	2	2	1	1	7
2	1	1	2	1	1	6
3	2	2	2	2	2	10
4	2	2	2	2	2	10
5	2	1	2	2	2	9
6	2	2	2	2	2	10
7	2	2	2	2	2	10
8	2	2	2	2	2	10
9	2	1	2	2	2	9
10	2	2	2	2	2	10
11	2	2	1	2	2	9
12	2	2	1	2	2	9
13	1	2	1	1	1	6
14	2	2	2	2	2	10
15	2	2	2	2	2	10

Trees 2 and 13 have the lowest parsimony score of 6.
Hence, these 2 trees is outputted (which is a tie) by MRP
for the input trees given in Figure 7.

Figure 10. Parsimony score of 4-taxon topologies.

It is easy to calculate the parsimony score and infer trees by hand where there are small numbers of input trees and taxa set. For large numbers of input trees, one can employ software packages to accomplish the task automatically.

3 Simulation study

The previous chapters provide some background information about this research. In this chapter, the simulation approach is discussed in detail, such as the procedure, possible pruning schemes, mutation model, usage of outgroup and the assessment.

The results for different known species trees with various settings are given at the end of this chapter to demonstrate the performance of MRP. We employ simulation in this study because the true species tree being known at the beginning makes it is easy to evaluate the performance of MRP for different conditions precisely and automatically.

3.1 Simulation procedure

For this simulation study, the gene trees sampled from a known species tree are used as the input trees and the output tree returned by MRP is treated as the estimated species tree. Using several packages and software, the simulation procedure involves the following 6 key steps, for which is developed and run under the Linux environment.

1. Sampling gene trees from a known true species tree,
   using COAL (Degnan and Salter, 2005);
2. Applying pruning scheme to the gene trees, using APE (Paradis et al., 2004);
3. Introducing mutation to the gene trees, if applicable

using Seq-Gen (Rambaut and Grassly, 1997) / PAUP* (Swofford, 2002);

4. Converting gene trees into matrix representation form,

using Clann (Creevey and McInerney, 2005);

5. Returning estimated species tree(s) based on the parsimony criterion,

using PAUP* (Swofford, 2002);

6. Assessing the performance,

using R (R Development Team, 2009) / PHANGORN (Klaus, 2010).

The procedure is carried out by repeating the first 5 steps $i = 300$ times such that for each iteration, a same number of gene trees is sampled firstly from the same species tree and then, the same combinations of mutation and pruning scheme are employed.

A different numbers of gene trees from the same species tree are sampled and run through the first 5 steps i times again with the same setting of mutation and pruning scheme, if needed. Finally, all the outputs from various numbers of gene trees are used to assess the performance of MRP for the species tree in the given settings of mutation and pruning scheme in the last step of the procedure.

It is efficient to apply pruning schemes before mutation when possible. This is because the sample size is reduced when pruning taxa from the gene trees. Therefore, it will take less time to complete the mutation afterwards. To this end, the simulation

procedure is run faster compared to the other way around.

3.2 Pruning schemes

3 main types of pruning schemes that can be applied to the gene trees are:

(i) randomly pruning a fixed number of taxa with equal probability of pruning each taxon; (ii) randomly pruning a different number of taxa with equal probability of pruning each taxon; and (iii) custom pruning schemes where randomly pruning a subset of possible taxa with unequal probability of pruning each taxon. The probability of pruning each individual taxon is non-negative and always adds up to 1 exactly for all these 3 cases.

The consensus setting is where each gene tree always has the same full set of taxa, which is equivalent to pruning no taxa. The supertree setting is used to address the case where some taxa are pruned from the gene trees, even when only 1 taxon is pruned. If a pruning scheme prunes all the possible taxa from every input gene tree, then no useful information is available, i.e. the gene trees are non-informative. Clearly, there are a maximal number of taxa that can be pruned from each gene tree so that every input gene tree is informative. Essentially, for the ith gene tree with $n_i \leq n$ taxa and the full taxon set has size $n = |S|$ taxa, one can prune at most $n_i - 3$ taxa from the ith gene tree; otherwise the ith gene tree is non-informative. For example, if the full taxon set size is 4, then at most $4 - 3 = 1$ taxon can be pruned from a given gene

tree. If more than 1 taxon is pruned, then the resulting gene tree is non-informative. In other words, the corresponding parsimony score of such gene tree is always the same against all the possible candidate trees.

It will be shown later that for some combinations of branch lengths and topologies the matching estimated species tree topology is returned less often with the consensus setting as the number of known gene trees increases. In contrast, the performance of MRP is improved with the supertree setting, for which the estimated matching species tree topology is more likely to be returned.

Because there are so many possible combinations of pruning schemes, we only explore some simple pruning schemes in the simulation to demonstrate the performance of MRP in these situations.

3.3 Mutation model

Introducing mutation to the simulated gene trees adds noise in order to obtain more realistic samples. The simplest 4-state mutation model, namely the Jukes-Cantor model (Jukes and Cantor, 1969; Felsenstein, 2004) is used to illustrate the mutation effect. In a nutshell, there are four bases A, G, C and T in a DNA sequence, A and G are called purines, C and T are called pyrimidines. In the Jukes-Cantor model, each base in the DNA sequence has the same probability to change within a given time. If 1

changes, it changes to 1 of the 3 remaining bases with the same probability. As a result, one can expect an equal frequency of the 4 bases DNA sequence after a certain amount of time units.

In this simulation study, for each individual gene tree, the DNA sequence are first generated by employing the software Seq-Gen and applying the Jukes-Cantor model for sequences of length 800 nucleotides. Then, gene trees are estimated from the resulting DNA sequences using neighbor-joining in PAUP*. If 2 or more estimated gene trees are outputted in a given iteration, only 1 is chosen at random. This is to ensure that the same numbers of gene trees are returned.

Neighbor-joining is used because it is a fast approximate distance method. Essentially, the clusters are used to estimate a tree from the distance (i.e. the branch lengths) matrix of taxa. The taxa (tips) with the lowest distance total between them are combined to form a cluster, which is then considered as a new taxon. This is repeated until 2 taxa are remained, and then attach the remaining 2 taxa directly by a branch length (Saitou and Nei, 1987; Felsenstein, 2004).

A key point to keep in mind is that the branch lengths of the species tree are measured in coalescent units, where a 1 unit branch length represents N_e generations, and N_e is the effective population size. However, the expected number of changes is used in the

mutation model, which is a different scale. The solution is to adjust the branch lengths of the simulated gene trees to the expected number of changes by multiplying the branch lengths with $\theta/2$ where $\theta = 0.01$ in Seq-Gen. This is because $\theta = 2N_e\mu$ and μ is the mutation rate per site per generation. Let T denote the number of generations. Then, we have that T/N_e is the corresponding branch length in coalescent time units, and $(\theta/2)\cdot T/N_e = \mu T$ is the expected number of changes that happen in T generations (Degnan and Rosenberg, 2009).

In this simulation study, we use the phrase "estimated gene trees" to refer to the situation where gene trees are obtained from the described mutation model above, and the phrase "simulated gene trees" when gene trees are purely simulated from the known true species tree.

3.4 Assessing the performance

The assessments are: (i) probabilities of inferring different species tree topologies and (ii) Robinson-Foulds distance (Robinson and Foulds, 1981). Both methods evaluate the consistency of MRP for inferring matching species trees in difference respects.

Method (i) is to calculate the proportion of times that each possible species tree is produced by MRP with the same settings. For any given iteration that yields k estimated species trees, each returned tree is weighted by $1 / k$. Then, for all iterations

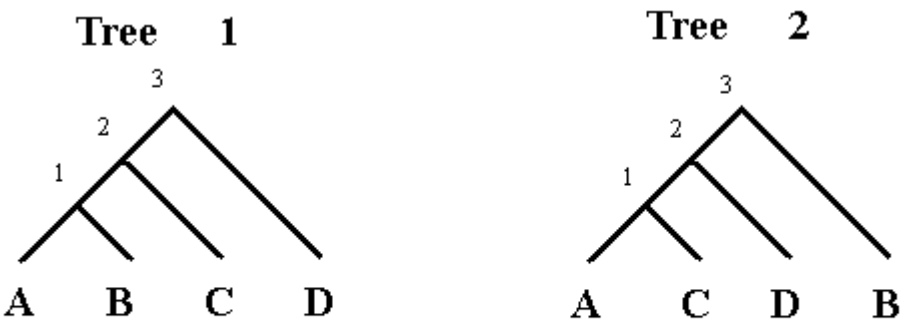
the weighted average is the overall proportion of times that each corresponding species tree is returned. To this end, method (i) gives the distribution of each estimated species tree outputted by MRP for particular settings. Consequently, MRP is statistically consistent if and only if the probability of returning the matching estimated species tree approaches 1, as the number of gene trees increases towards infinity.

Method (i) is useful for species tree with a small number taxa, i.e. less than 10, because the number of possible candidate trees increases dramatically as the number of taxa increases. For example, there are more than 30 million possible trees with 10 taxa (Felsenstein, 2004), making it hard to cover all of them in the simulation.

On the other hand, method (ii) is fast especially for trees with many taxa, e.g. 10 and more. This is because only the estimated species tree returned by MRP and the true species tree are used to calculate the Robinson-Foulds distance.

Method (ii) is defined as the following in this simulation study. For only 2 trees A and B, let a be the number of clades that occur in tree A but not in tree B, similarly let b be the number of clades that occur in tree B but not in tree A. Then the sum, $d = a + b$ is the Robinson-Foulds distance of these 2 trees. For iterations that output more than 1 estimated species tree, d is calculated for each estimated species tree against the true

species tree separately and the average is used as the Robinson-Foulds distance for that iteration. The overall Robinson-Foulds distance *D* is given by finding the average of *d*. Figure 11 is an example of the Robinson-Foulds distance.



Trees 1 and 2 have 2 clades that do not exist in the other one (excluding the node 3). Hence the Robinson-Foulds distance is $2 + 2 = 4$ in this example.

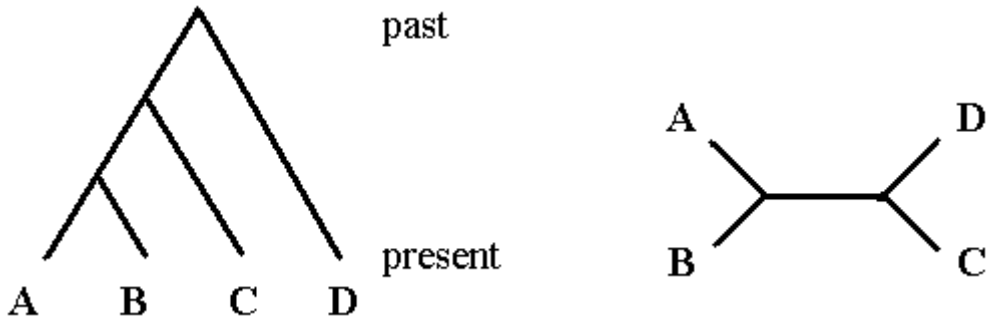
Figure 11. Robinson-Foulds distance.

There are $n - 1$ internal nodes for a bifurcating tree with *n* taxa. With the same taxa set, the maximal nodes that 2 bifurcating trees can not agree are $(n - 2)$, as both of them always have the same clade (i.e. the full taxa set) at the root. Thus, the maximal Robinson-Foulds distance of 2 bifurcating trees with the same *n* taxa is $(n - 2) \times 2$. The Robinson-Foulds distance measurement is normalized such that it is a score between 0 and 1 by dividing with the maximal Robinson-Foulds distance. Then, a normalized Robinson-Foulds distance with a score of 0 indicates that the estimated species tree topology is exactly same as the true species tree topology. Similarly, a score of 1 means these 2 topologies are completely different from each other. Therefore, the estimated species tree topology is more similar to the true species tree topology if the corresponding normalized Robinson-Foulds distance is closer to 0 and more dissimilar if the score is near 1. One can say MRP is statistically consistent if

and only if the respective normalized Robinson-Foulds distance is approaching 0, as the number of gene trees increases.

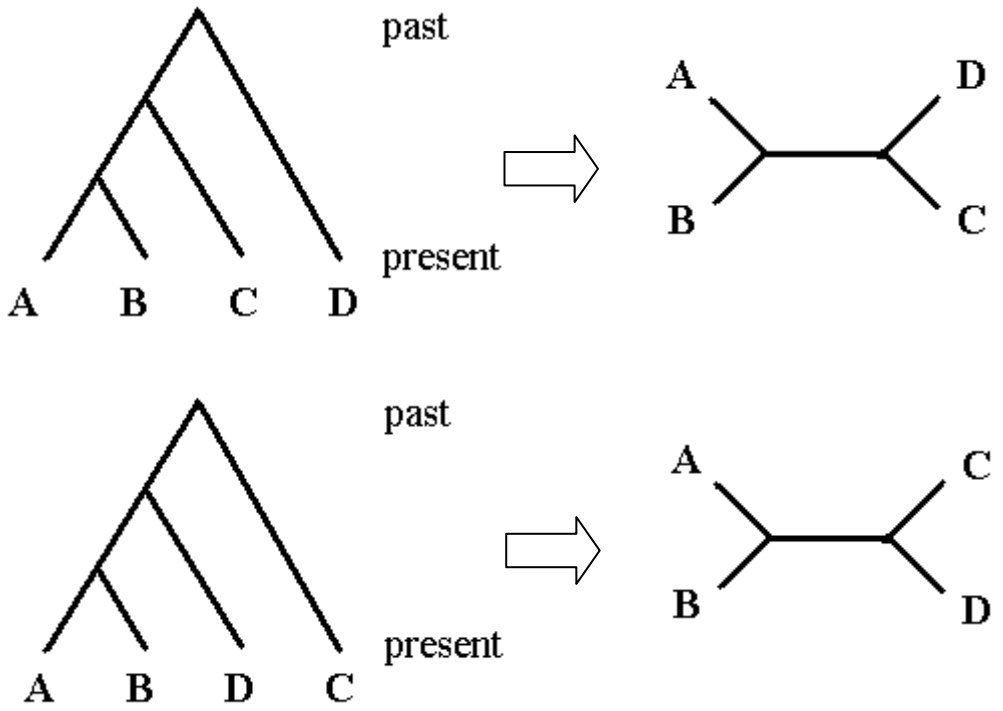
3.5 Outgroup

Both species trees and gene trees can be in rooted or unrooted topology form. The main difference is that in the unrooted case (Figure 12), there is no direction of time. Because of this, it is possible for species trees and gene trees with different rooted topologies to end up having the same unrooted topology (Figure 13). Consequently, it is hard to distinguish different topologies of species trees and gene trees in such a situation.



Rooted Unrooted

Figure 12. Rooted and unrooted topologies.



Rooted Unrooted

The top and bottom left rooted trees have the same unrooted topology because both trees have the same clade {AB} at the left side and clade {CD} at the right, although these 2 trees have different rooted topologies.

Figure 13. Same unrooted topology.

On the other hand, one can recover a rooted topology from an unrooted topology by rooting at a selected taxon. Notice that it is possible to produce different rooted topologies by rooting at different taxa even starting from the same unrooted topology, as illustrated in Figure 14.

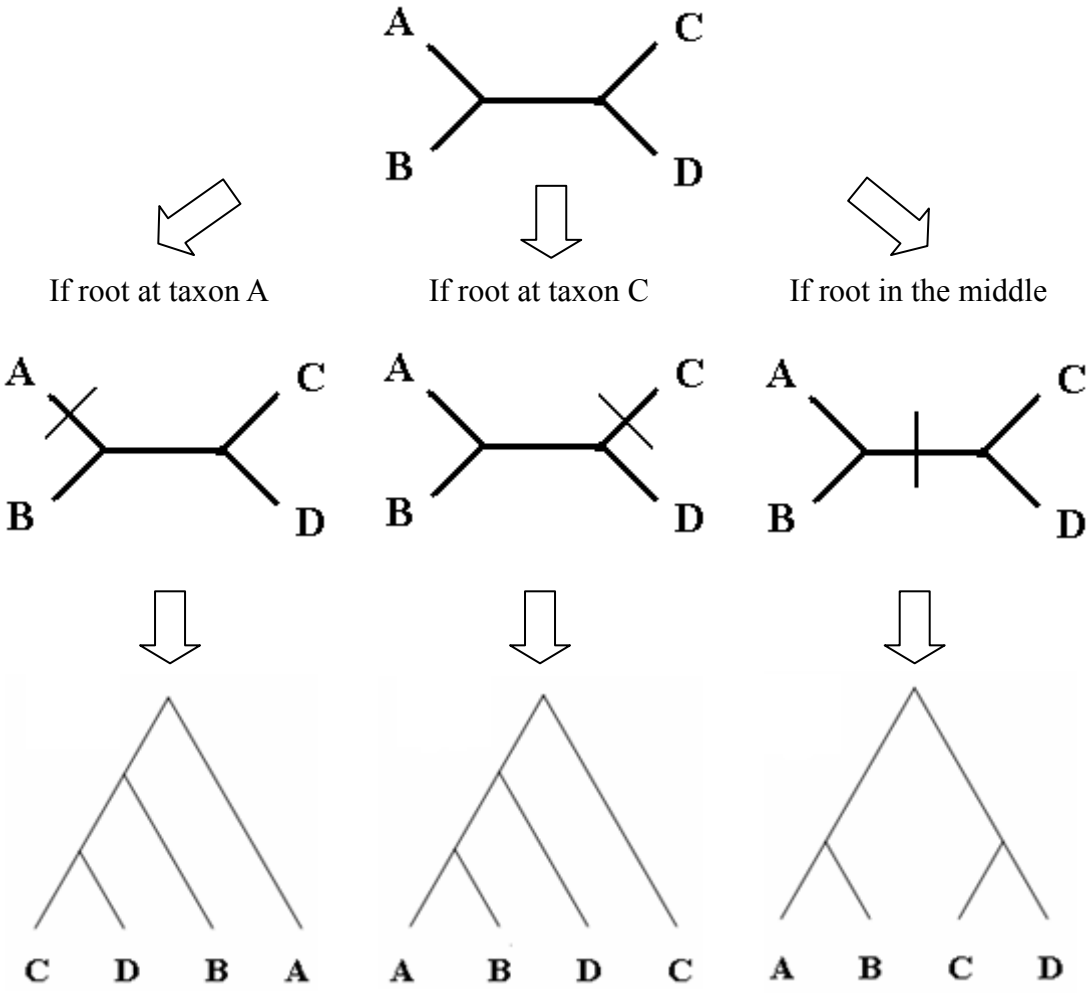
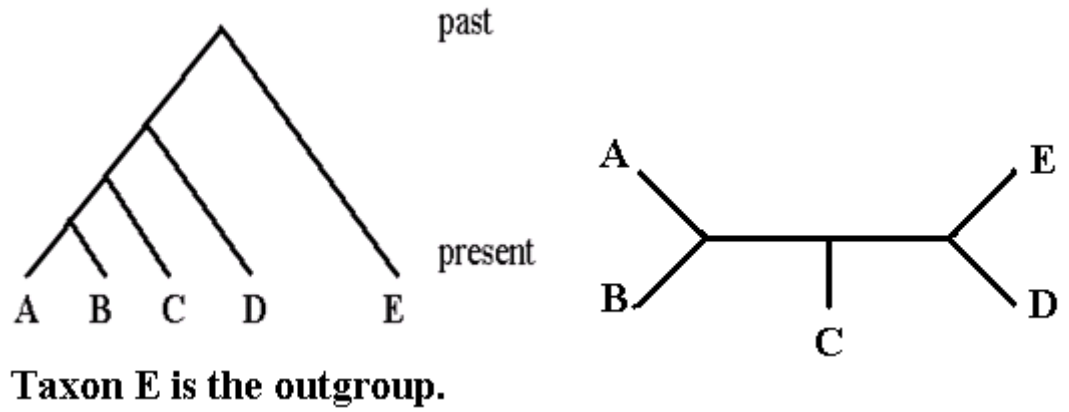


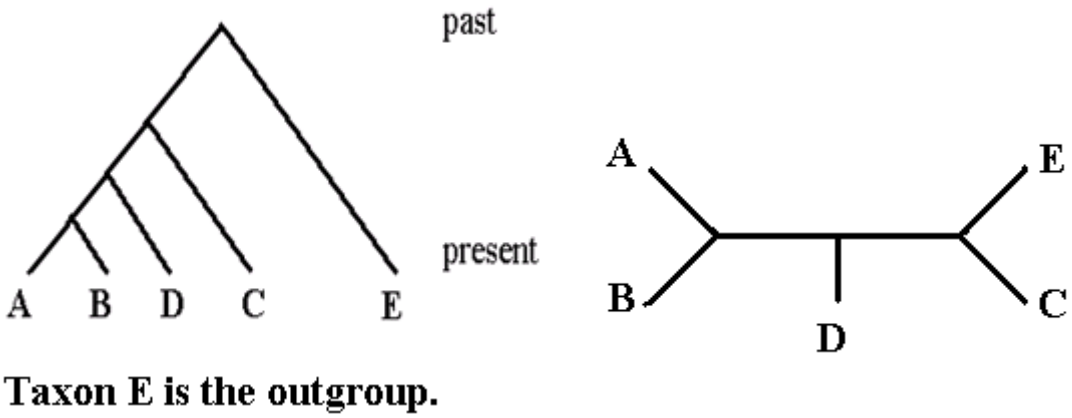
Figure 14. Rooting the unrooted topology.

For this simulation study, both the estimated gene trees and the estimated species tree are outputted in unrooted topologies from PAUP*. However, both the true species tree and the corresponding simulated gene trees are in rooted topologies. In order to avoid the issue illustrated in Figures 13 and 14, a taxon *X* is used as the outgroup with a large branch length (e.g. 30 coalescent units) attached to the rooted true species tree such that all simulated gene trees from this species tree are automatically rooted and have taxon *X* as the outgroup. Taxon *X* is not deleted from any of the gene trees with any pruning schemes. For the case of mutation, all estimated gene trees are always rooted at taxon *X*. It is essential to root all trees at the same outgroup, otherwise

unreliable gene trees will produce. That is, all gene trees and species tree will always have the same taxon *X* as the outgroup within the simulation procedure to preserve the topologies.



Taxon E is the outgroup.



Taxon E is the outgroup.

Rooted Unrooted

Different unrooted topologies from different rooted topologies because trees have different clades at the right side, e.g. {DE} and {CE} respectively.
If rooted the unrooted topologies at taxon E will always return the corresponding rooted topologies to the left.

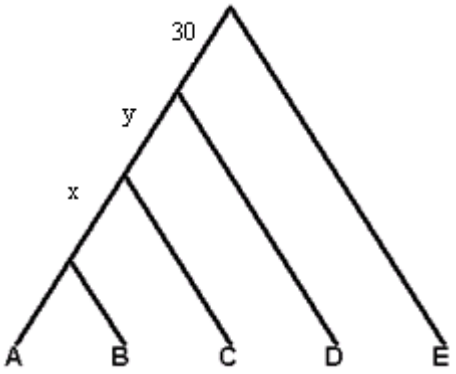
Figure 15. Outgroup effect example.

3.6 Species trees with 4 taxa and outgroup

In this setting, topologies of 4-taxon ((((A,B),C),D),E) and (((A,B),(C,D)),E) with outgroup taxon E were examined. The performance of MRP was tested for selected branch lengths with different combinations of mutation and pruning. The pruning scheme was either always randomly deleting 0 or 1 taxon from each gene tree used so that all of them are informative. All the possible resulting settings were carried out by following the described simulation procedure above for various number of simulated and estimated gene trees: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2500 and 3000 with 300 replications in each case. The assessment was to calculate the probabilities of returning different estimated species tree topologies. Relevant results are discussed to show the performance of MRP under different circumstances.

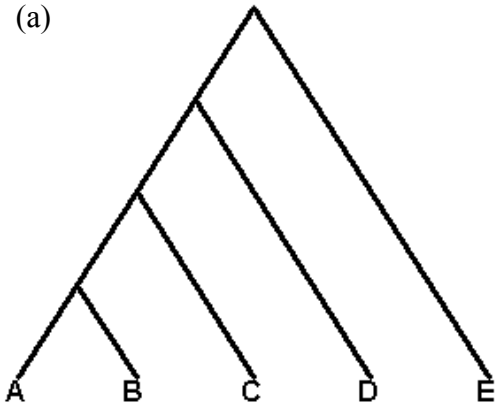
3.6.1 Under topology ((((A,B),C),D),E)

Figure 16 illustrates the true species tree model used and a possible MRP estimated species tree topology. There were 2 sets of branch lengths measured in coalescent units used, (x, y) = (0.05, 0.05) and (0.1, 0.1). Hence, there were 8 possible cases: 2 sets of branch lengths, 2 pruning schemes and 2 kinds of gene trees (simulated and estimated). Only 4 of the most frequently returned estimated species trees topology were reported to show the performance of MRP in different settings.



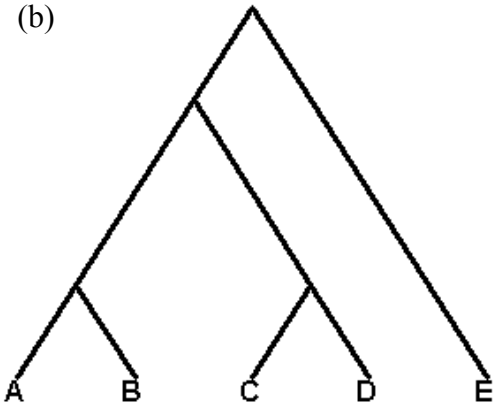
The species tree used in the simulation with branch lengths (x, y)
and a taxon E as the outgroup with branch length of 30 coalescent units.

(a)



The matching MRP estimated species tree topology with the same outgroup.

(b)



A non-matching MRP estimated species tree topology with the same outgroup.

Figure 16. Species tree model and possible outcomes.

The results show the effects of branch lengths, mutation and pruning scheme. The non-matching estimated species tree topology in Figure 16b was returned most frequently, when the branch lengths (x, y) were short, e.g. (x, y) = (0.05, 0.05). With these branch lengths, a larger proportion of simulated gene trees from this species tree had the non-matching topology (Figure 16b) than the matching topology (Figure 16a). The non-matching estimated species tree topology Figure 16b was more often returned from MRP, regardless of the number of gene trees (top left of Figure 17).

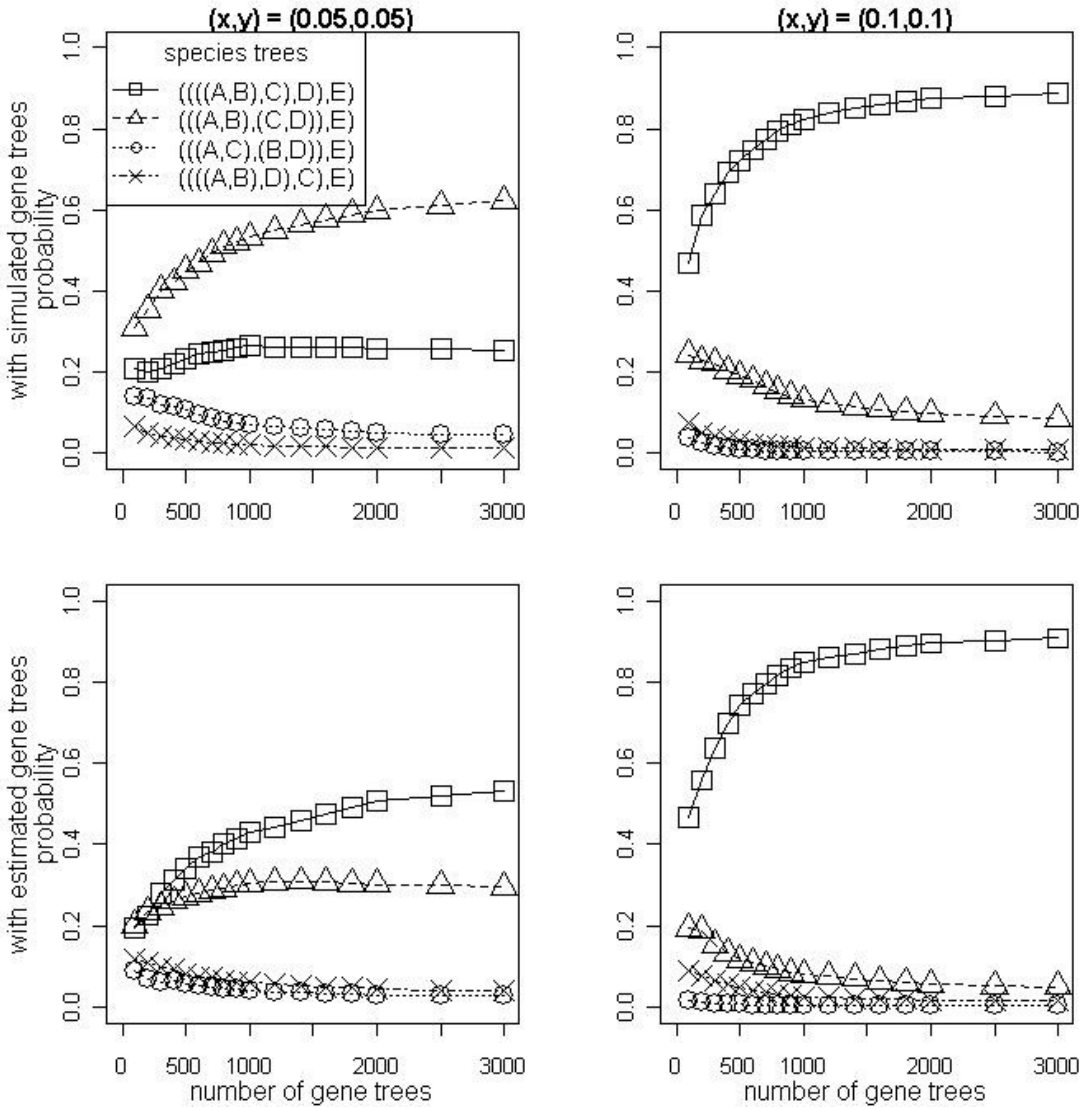


Figure 17. No pruning under the true species tree in Figure 16. All plots have the same legend as the top left one.

If the most frequently outputted estimated species tree topology matched with the true species tree topology, then the proportions of time of returning the matching species trees was significantly lower with shorter branch lengths, (x, y) = (0.05, 0.05) compared to (0.1, 0.1). For example, about 37 % less often with 3000 gene trees, respectively (compare the bottom row of Figure 17).

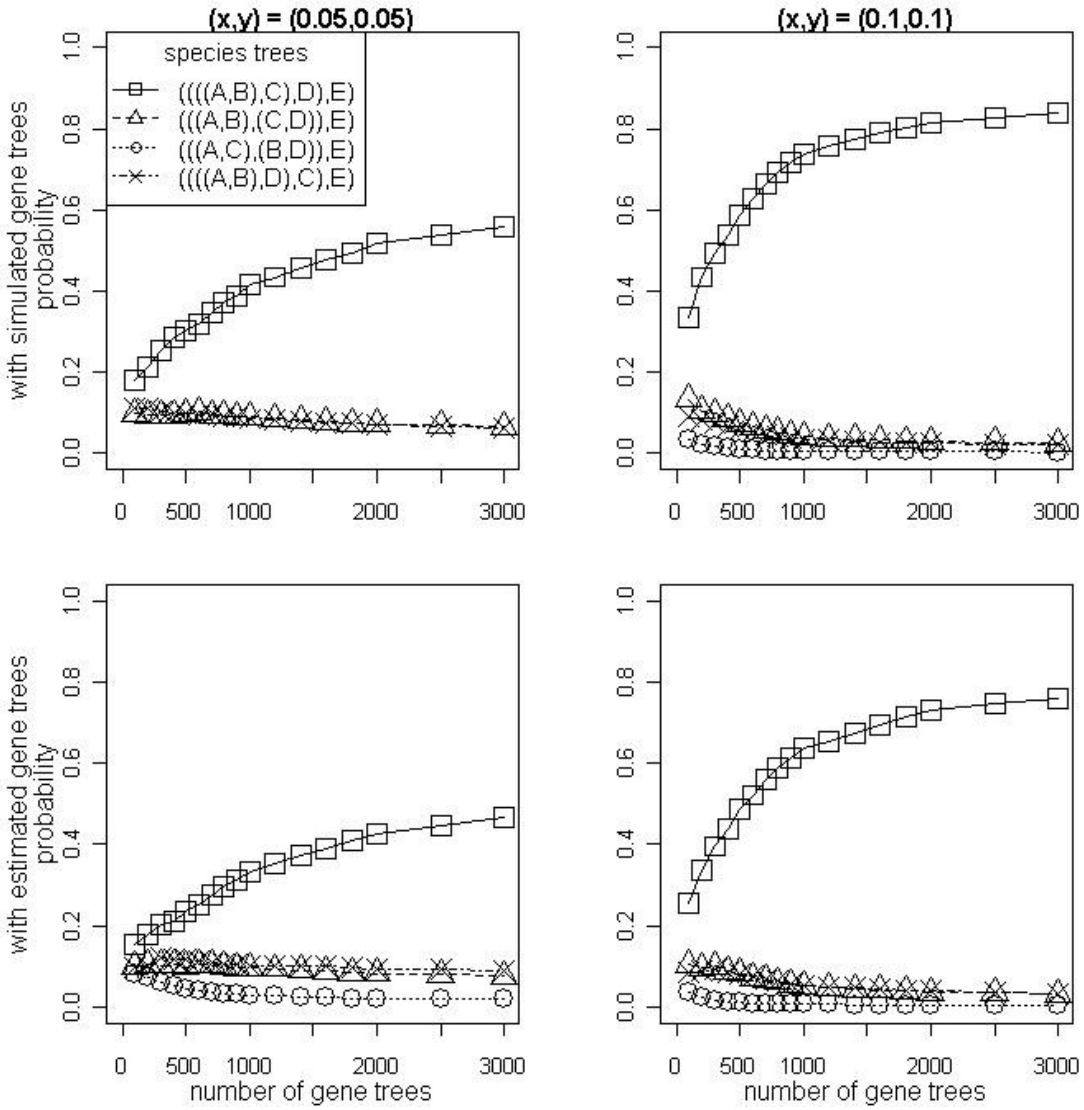


Figure 18. Pruning 1 taxon randomly under the true species tree in Figure 16. All plots have the same legend as the top left one.

The performance of MRP was affected by applying the pruning scheme to the gene trees. For instance, the most probable MRP estimated tree topology matched the true species tree topology with a short branch lengths $(x, y) = (0.05,0.05)$ when exactly 1 taxon was pruned from the simulated gene trees. However, when no pruning was used the most frequent MRP estimated species tree topology did not match the true species tree topology (compare the top left of Figures 17 and 18). Because there was less information available with pruned gene trees, this result suggested that sometimes the

performance of MRP was improved with less information. This counterintuitive result can be explained by the robustness of supertree methods using rooted triples (Wilkinson et al., 2005; Steel and Rodrigo, 2008). It follows that by randomly pruning 1 taxon, all the gene trees were in the form of rooted triple. This suggests that MRP is robust with such gene trees in the sense that the matching estimated species tree topology is returned most often.

However, applying pruning schemes can diminish the performance of MRP in some cases. For example, with 3000 simulated gene trees with the branch lengths $(x, y) = (0.1,0.1)$, the matching estimated species tree topology was returned at about 88.9% of the time without pruning scheme; but only 83.8% if 1 taxon was pruned from each gene tree (compare the top right of Figures 17 and 18). Similar patterns can be found in the rest respective pairs of the figures. Hence, under the true species tree topology ((((A,B),C),D),E), these results suggested that there was no unique effect of the pruning scheme on the performance of MRP.

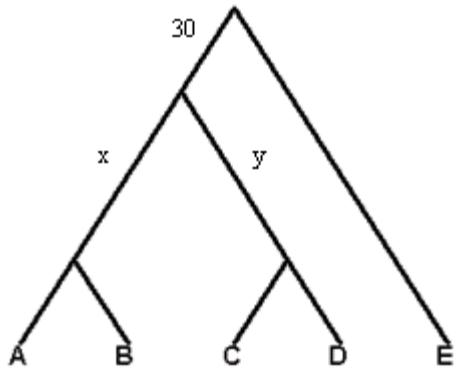
The performance of MRP with estimated gene trees had similar outcomes as the pruning schemes under the true species tree topology ((((A,B),C),D),E) in the simulation. On the one hand, the matching estimated species tree topology was yielded more often if using estimated gene trees with no pruning (28 % and 2 % more often for 3000 estimated gene tree, as shown in the top and bottom rows of Figure 17).

On the other hand, the matching estimated species tree topology was returned less frequently if using estimated gene trees with pruning 1 taxon (about 9% and 7.5% drop for 3000 estimated gene trees, compare the top and bottom rows of Figure 18).

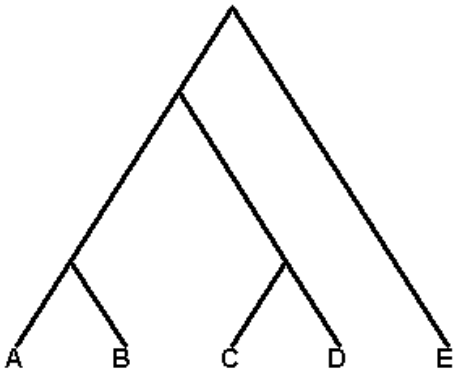
3.6.2 Under topology (((A,B),(C,D)),E)

The underlying true species tree model is given below with the branch lengths (x, y) and the matching MRP estimated species tree topology (Figure 19). The same sets of branch lengths measured in coalescent units as topology ((((A,B),C),D),E) were used. Namely, (x, y) = (0.05, 0.05) and (0.1, 0.1). This group of branch lengths makes it more convenient to compare the results between topologies (((A,B),(C,D)),E) and ((((A,B),C),D),E) as the same 8 cases are covered. The results presented the top 3 most frequently estimated species trees topology to show the performance of MRP.

Although the most frequently estimated MRP species tree topology did always match the true species tree topology, it was yielded less often with shorter branch lengths (compare the left and right columns of Figures 20 and 21). If shorter branch lengths were used, e.g. (x, y) = (0.05, 0.05), the corresponding gene trees sampled from resulting species tree were still more likely to have the matching topology than any other topology. It was therefore not supervising that the performance of MRP for true species tree topology (((A,B),(C,D)),E) was less affected by the branch lengths than for topology ((((A,B),C),D),E) in the simulation.



The species tree used in the simulation with branch lengths (x, y)
and a taxon E as the outgroup with branch length of 30 coalescent units.



The matching MRP estimated species tree topology with the same outgroup.

Figure 19. Species tree model and possible outcomes.

The matching MRP estimated species tree topology was returned less frequently when pruning 1 taxon from the gene trees. For example, using 3000 simulated gene trees for branch lengths (x, y) = (0.05, 0.05) and (0.1, 0.1), the matching estimated species tree topology was yielded about 32% and 10% less often, respectively (compare the top rows of Figures 20 and 21).

Similarly, with estimated gene trees, the matching estimated species tree topology was returned less often. For instance, with 3000 estimated gene trees and when no pruning was used, the matching estimated species tree topology was produced about 5% and 2% less frequently for the corresponding branch lengths (x, y) = (0.05, 0.05) and

(0.1, 0.1), comparing the top and bottom rows of Figure 20. In the same way, if using 3000 estimated gene trees and also 1 taxon was randomly pruned, the matching estimated species tree topology was outputted about 15% and 8% less often, respectively (compare the top and bottom rows of Figure 21).

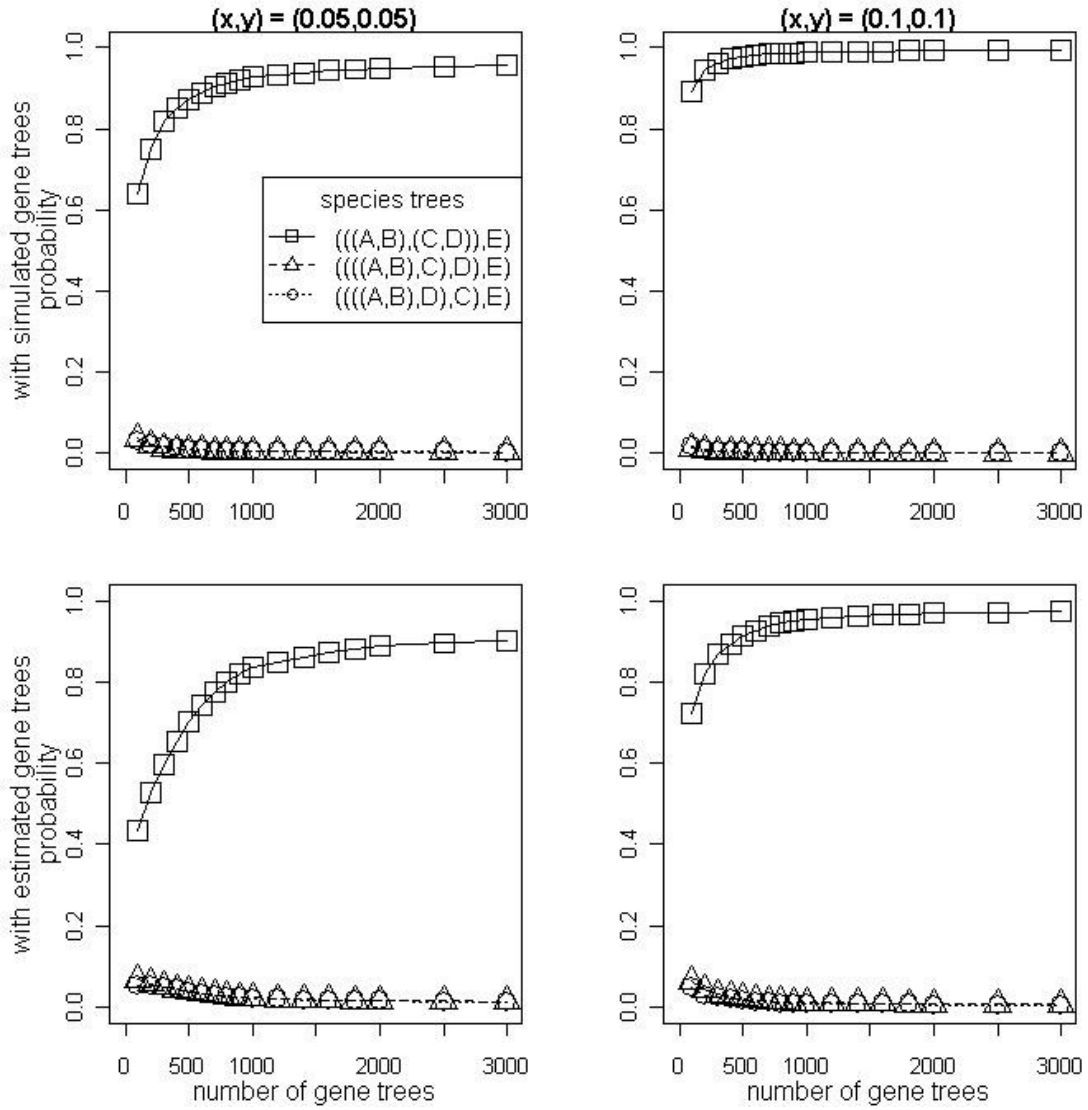


Figure 20. No pruning under the true species tree in Figure 19. All plots have the same legend as the top left one.

In summary, the results of both topologies ((((A,B),C),D),E) and (((A,B),(C,D)),E) in the simulation suggested that the MRP estimated species tree topology matched the true species tree topology more often with longer branch lengths, e.g. (x, y) = (0.1, 0.1)

compared to (0.05, 0.05). Applying pruning schemes and using estimated gene trees can either improve or reduce the proportion of times that MRP returns the matching estimated species tree topology, but this depends on the branch lengths and the true species tree topology. Nevertheless, only some cases of the 4-taxon true species tree with outgroup were used in the simulation to give a sense of the MRP performance.

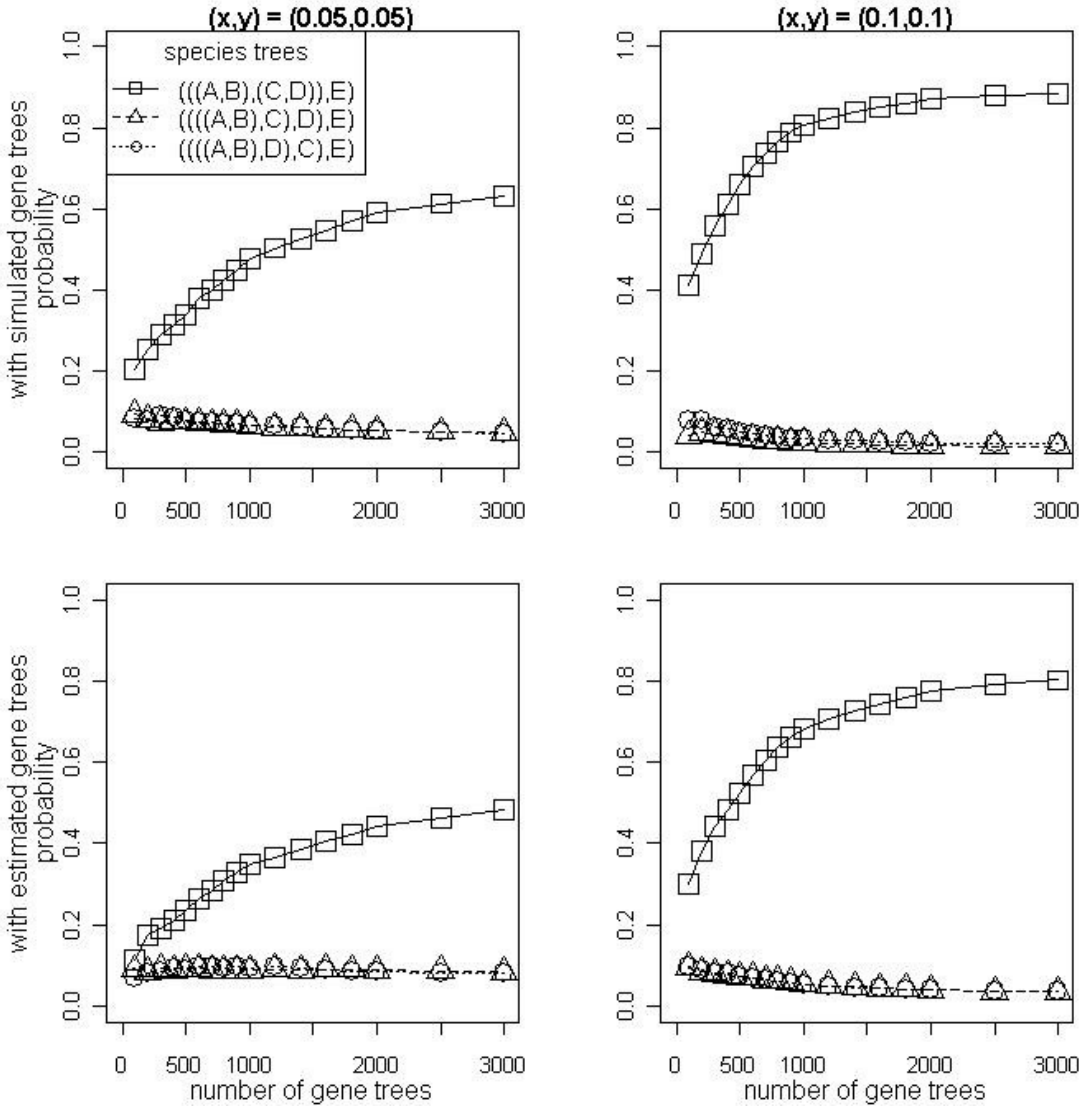


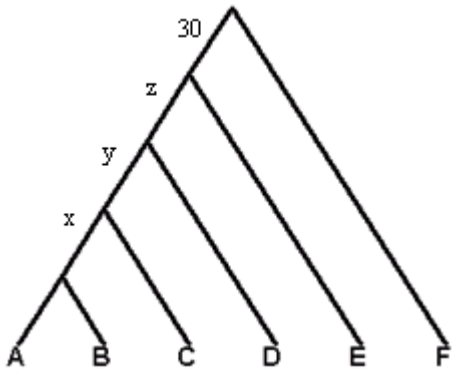
Figure 21. Pruning 1 taxon randomly under the true species tree in Figure 19. All plots have the same legend as the top left one.

3.7 Species trees with 5 taxa and outgroup

3 topologies of 5-taxon were examined: (((((A,B),C),D),E),F), ((((A,B),(C,D),E),F) and ((((A,B),C),(D,E)),F) with outgroup taxon F for selected branch lengths with various settings to assess the performance of MRP. The pruning scheme deleted 0, 1 or 2 taxa from each gene tree randomly so that all the gene trees were informative. All the possible combinations of selected branch lengths and pruning schemes were conducted by following the described simulation procedure (Section 3.1) for various numbers of simulated and estimated gene trees: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2500 and 3000 with 300 replications in each case, respectively. The assessment was the probability of inferring the topologies of each candidate species tree. Relevant results were reported and discussed to show the performance of MRP under different conditions.

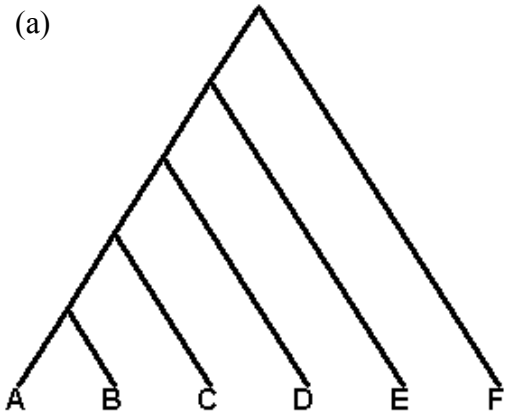
3.7.1 Under topology (((((A,B),C),D),E),F)

The true species tree model is given with the branch lengths (x, y, z), and the possible MRP estimated species tree topologies (Figure 22). The selected sets of branch lengths measured in coalescent units (x, y, z) = (0.1, 0.1, 0.1), (2.0, 0.05, 0.05) and (0.05, 0.05, 2.0) were used. Hence, there are 18 possible cases, 3 set of branch lengths, 3 pruning schemes and 2 types of gene trees. Only 3 estimated species trees were reported in each case with discussion to demonstrate the performance of MRP.



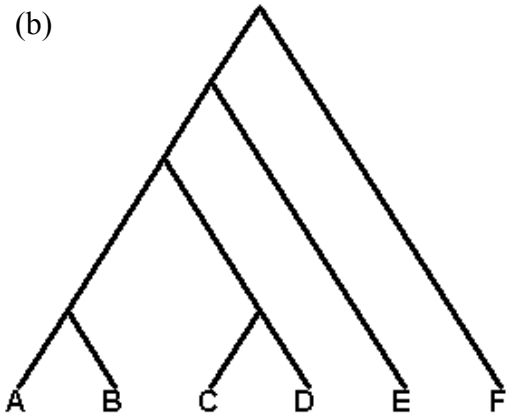
The species tree used in the simulation with branch lengths (x, y, z)
and a taxon F as the outgroup with branch length of 30 coalescent units.

(a)

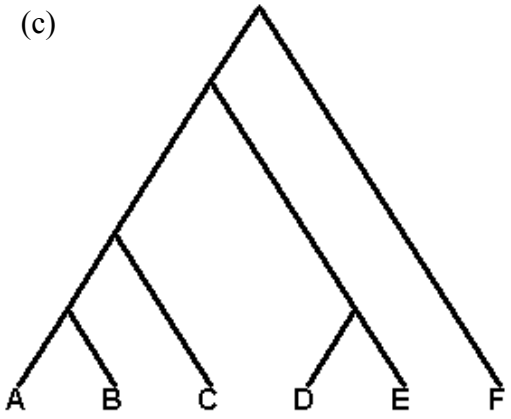


The matching MRP estimated species tree
topology with the same outgroup.

(b)



(c)



Two non-matching MRP estimated species tree topologies with the same outgroup.

Figure 22. Species tree model and possible outcomes.

The results illustrate that the most frequently returned estimated species tree topology was different for these 3 sets of branch lengths with 3000 simulated gene trees (compare the top row of Figure 23). For instance, with branch lengths (x, y, z) =

(0.1, 0.1, 0.1), the most frequently outputted estimated species tree topology did not match the true species tree topology for small numbers of gene trees. However, the matching estimated species tree topology was returned more often with larger numbers of simulated gene trees when there was more information available to identify the matching estimated species tree topology (top left of Figure 23).

The performance of MRP was about the same for (x, y, z) = (2.0, 0.05, 0.05) and (0.05, 0.05, 2.0), see the top middle and right of Figure 23. In both cases, the non-matching estimated species tree topology (Figure 22b and c) was yielded more often regardless of the number of simulated gene trees.

In addition, the matching estimated species tree topology was produced more often with the branch lengths (x, y, z) = (0.1, 0.1, 0.1) than with (2.0, 0.05, 0.05) and (0.05, 0.05, 2.0) under the same simulation settings, (compare the left columns against the other 2 columns of Figures 24 and 25). The branch lengths total of (x, y, z) = (0.1, 0.1, 0.1) is 0.3 coalescent units, which is smaller but more evenly distributed than the other 2 sets of branch lengths, (x, y, z) = (2.0, 0.05, 0.05) and (0.05, 0.05, 2.0). This observation suggests that the matching estimated species tree topology (((((A,B),C),D),E),F) was returned more often with evenly distributed internal branch lengths, like (x, y, z) = (0.1, 0.1, 0.1) in the simulation.

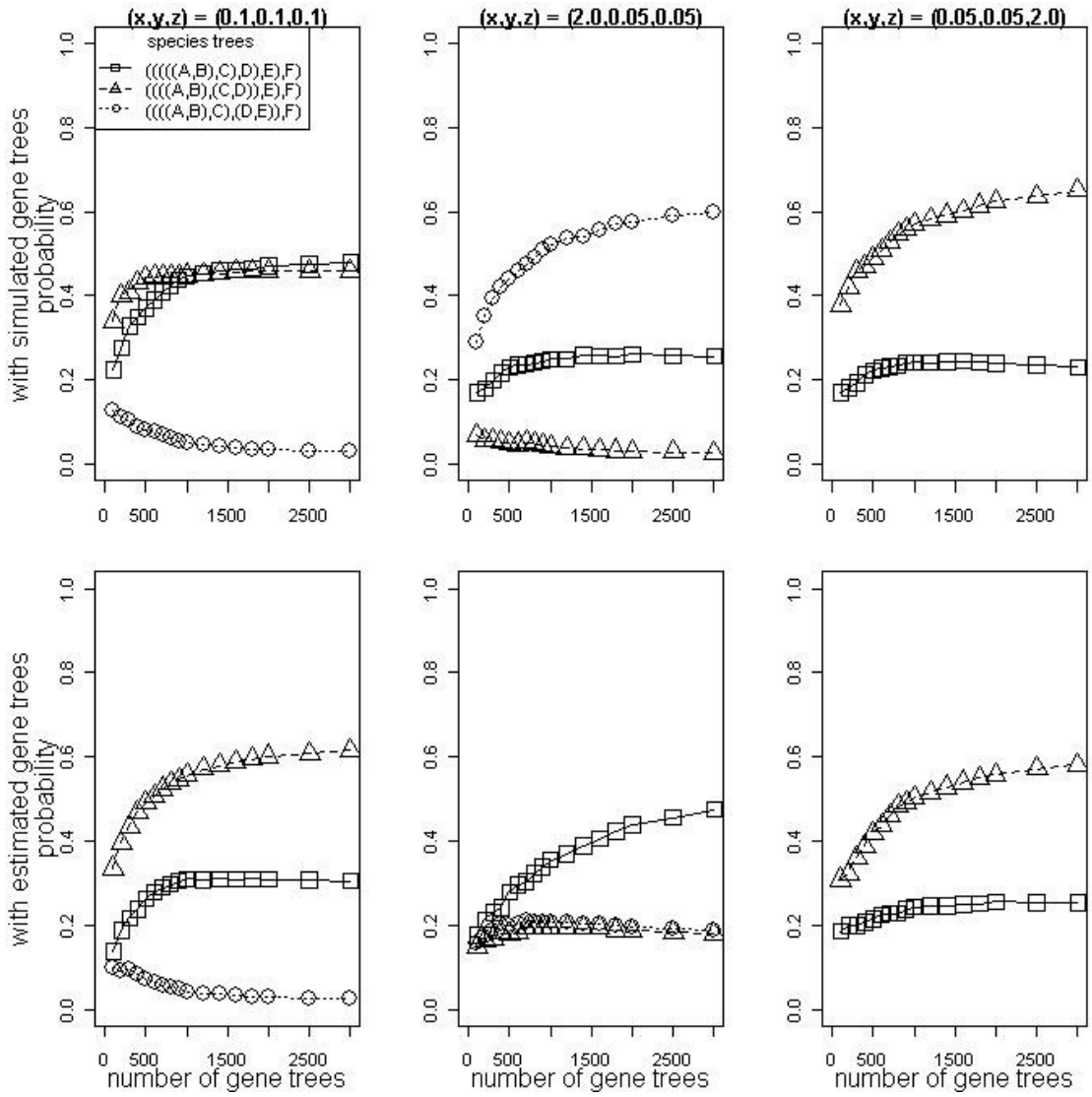


Figure 23. No pruning under the true species tree in Figure 22. All plots have the same legend as the top left one.

The matching estimated species tree topology (((((A,B),C),D),E),F) was returned more frequently by MRP when at least 1 taxon was randomly pruned from the gene trees used. This was true for all the selected branch lengths sets in the simulation: (x, y, z) = (0.1, 0.1, 0.1), (2.0, 0.05, 0.05) and (0.05, 0.05, 2.0). However, the matching estimated species tree topology was yielded less often if 2 taxa were randomly pruned compared to 1 taxon from the input gene trees in some cases. For example, the proportion of time for returning the matching species tree topology was

reduced by about 30%, 17% and 25%, if using estimated gene trees with 1 more taxon pruned (compare the bottom rows of Figures 24 and 25). This occurs because when gene trees had more taxa pruned, less information was available so that a larger sample of gene trees were needed to find the matching estimated gene trees topology. This observation also can be found with 4-taxon species tree topologies from the simulation (Section 3.6).

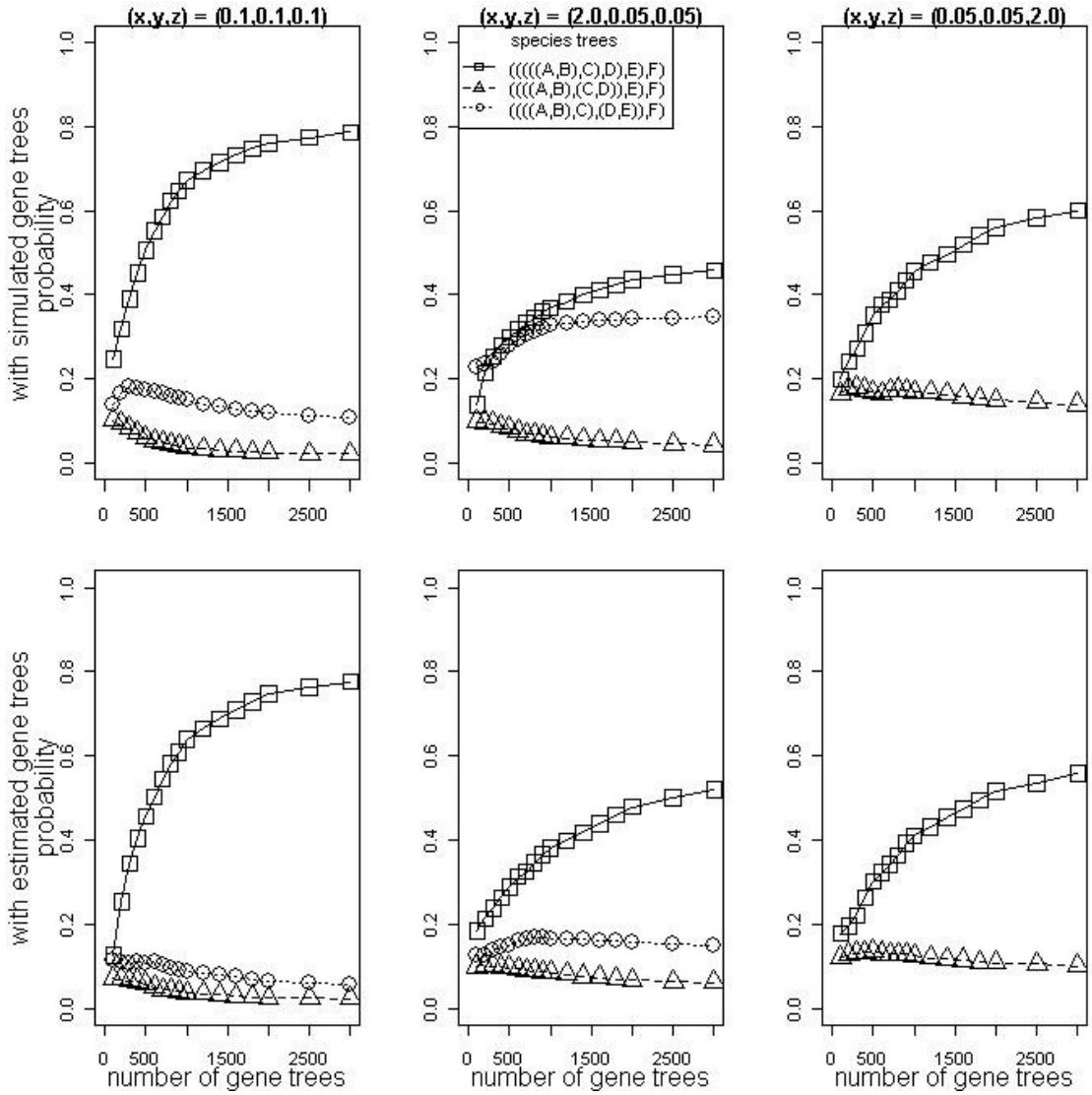


Figure 24. Pruning 1 taxon randomly under the true species tree in Figure 22. All plots have the same legend as the top middle one.

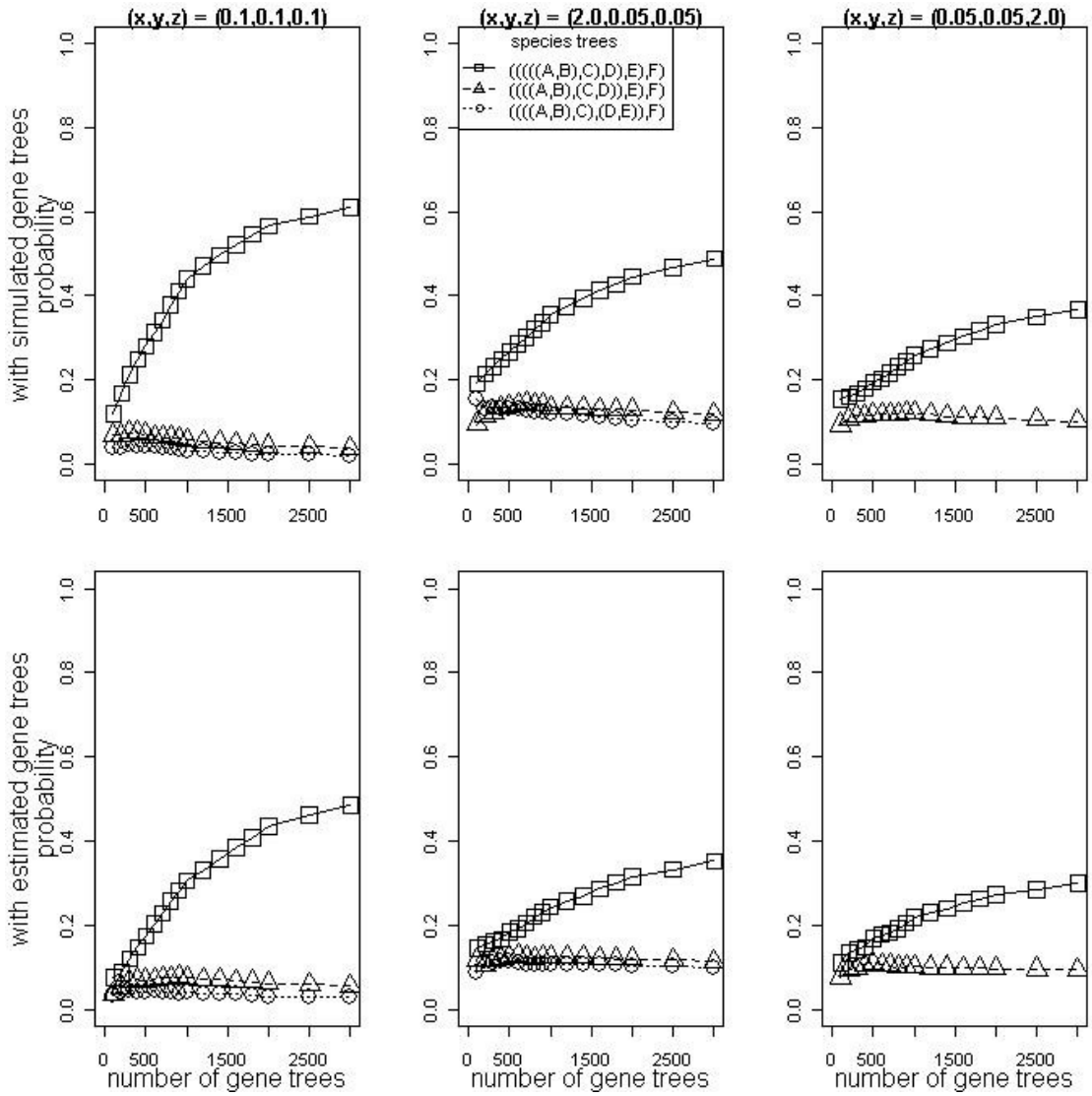


Figure 25. Pruning 2 taxa randomly under the true species tree in Figure 19. All plots have the same legend as the top middle one.

The performance of MRP using estimated gene trees from DNA sequences depended on the branch lengths (x, y, z) when no pruning was used (Figure 23). For (x, y, z) = (0.1, 0.1, 0.1), a non-matching estimated species tree topology was returned more often (left column of Figure 23). For example, the matching estimated species tree topology (((((A,B),C),D),E),F) was yielded about 17% less often, when using 3000 estimated gene trees instead of simulated gene trees. In contrast, the most frequently estimated species tree topology returned by MRP matched the true species

tree topology (middle column of Figure 23) with branch lengths (x, y, z) = (2.0, 0.05, 0.05). However, using estimated gene trees with (x, y, z) = (0.05, 0.05, 2.0), there was no significant effect on the performance of MRP. It was about the same proportion of time to yield both of the matching and non-matching estimated species trees topologies with either simulated or estimated gene trees, as shown in the right column of Figure 23.

With estimated gene trees, the performance of MRP was changed when pruning 1 taxon instead of no pruning. The matching estimated species tree topology (((((A,B),C),D),E),F) was returned about the same proportion of time for branch lengths (x, y, z) = (0.1, 0.1, 0.1) for both estimated and simulated gene trees (left column of Figure 24). However, it was about 6% more often to output the matching estimated species tree topology for (x, y, z) = (2.0, 0.05, 0.05), and about 5% less often for (x, y, z) = (0.05, 0.05, 2.0), when 1 taxon was randomly pruned with 3000 estimated rather that simulated gene trees (middle and right columns of Figure 24). Nevertheless, when 2 taxa were randomly pruned, the matching estimated species tree topology was returned less often if using estimated gene trees (compare the top and bottom rows of Figure 25).

In conclusion, the effect of using estimated gene trees on the performance of MRP was related to the branch lengths used in the simulation under the true species tree

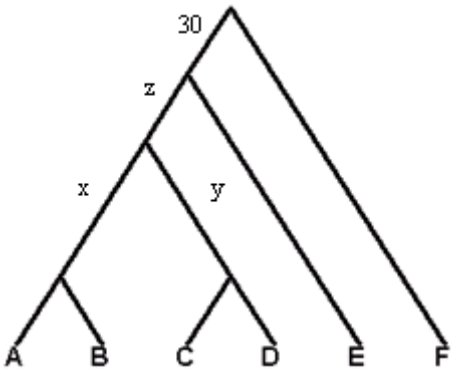
topology, (((((A,B),C),D),E),F). The matching estimated species topology was outputted more often when at least 1 taxon was pruned and using evenly distributed branch lengths in the simulation for both simulated and estimated gene trees.

3.7.2 Under topology ((((A,B),(C,D),E),F)

All gene trees were sampled from the true species tree model (Figure 26) with the branch lengths (x, y, z), and the possible MRP estimated species tree topologies. The following 3 sets of branch lengths measured in coalescent units were used: (x, y, z) = (0.1, 0.1, 0.1), (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05). There are 18 cases: 3 sets of branch lengths; 3 pruning schemes, and 2 types of gene trees (simulated and estimated). Only 3 estimated species trees were reported to demonstrate the performance of MRP.

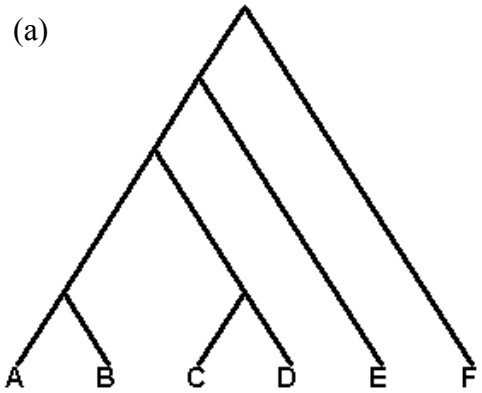
The matching estimated species tree topology ((((A,B),(C,D),E),F) was returned more frequently with the uniformly distributed branch lengths (x, y, z) = (0.1, 0.1, 0.1) than with the other 2 sets of branch lengths under the same conditions from the simulation (left columns of Figures 27 – 29). In addition, the non-matching estimated species tree topologies, (((A,B),((C,D),E)),F) (Figure 26c) and ((((A,B),E),(C,D)),F) (Figure 26b) were yielded with about the same frequency with branch lengths (x, y, z) = (0.1, 0.1, 0.1). For both branch lengths (x, y, z) = (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05) in the same simulation settings, the matching estimated species tree topology

((((A,B),(C,D),E),F) was yielded roughly the same proportion of the time (compare the middle and right columns of Figures 27 – 29).



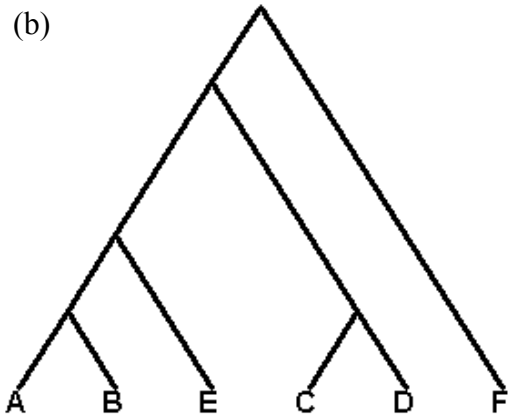
The species tree used in the simulation with branch lengths (x, y, z) and a taxon F as the outgroup with branch length of 30 coalescent units.

(a)

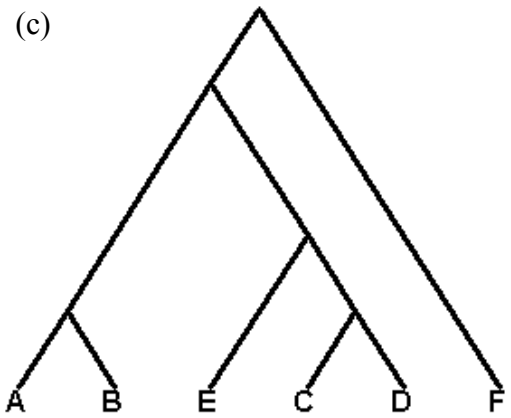


The matching MRP estimated species tree topology with the same outgroup.

(b)



(c)



Two non-matching MRP estimated species tree topologies with the same outgroup.

Figure 26. Species tree model and possible outcomes.

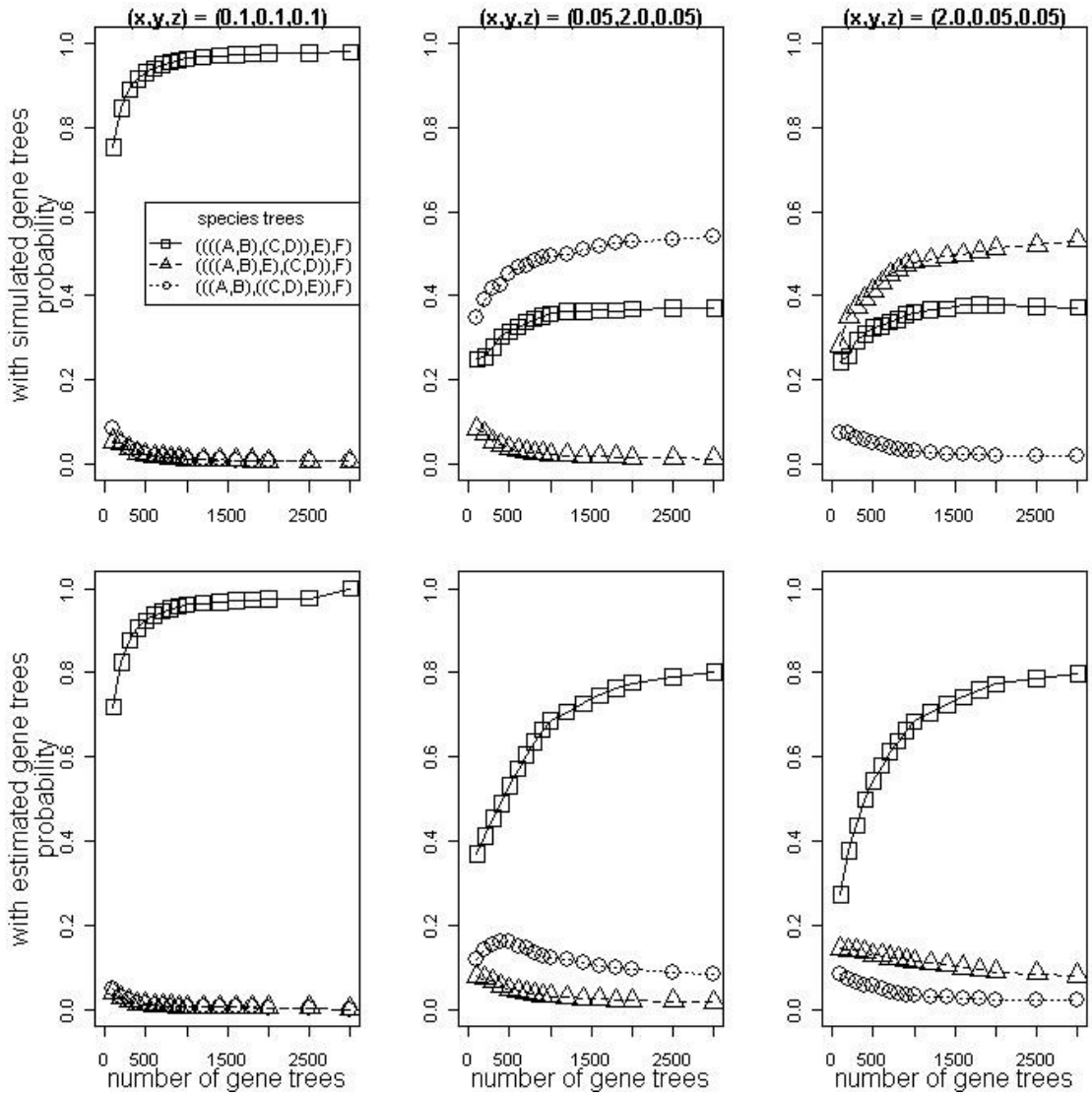


Figure 27. No pruning under the true species tree in Figure 26. All plots have the same legend as the top left one.

The non-matching estimated species tree topologies (((A,B),((C,D),E)),F) and ((((A,B),E),(C,D)),F) were returned most often with the same proportion of time, for the corresponding branch lengths (x, y, z) = (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05), if using simulated gene trees (compare the top row of Figure 27). The non-matching estimated species tree topology, (((A,B),((C,D),E)),F) was produced more often than ((((A,B),E),(C,D)),F) if using branch lengths (x, y, z) = (0.05, 2.0, 0.05). In contrast, if using (x, y, z) = (2.0, 0.05, 0.05) instead, it was the other way around with about the

same frequency (compare the middle and right columns of Figures 27 – 29). Nevertheless, both {AB} and {CD} clades were contained in the most frequently returned estimated species tree topologies, ((((A,B),(C,D),E),F), (((A,B),((C,D),E)),F) and ((((A,B),E),(C,D)),F) for all the selected branch lengths in the simulation.

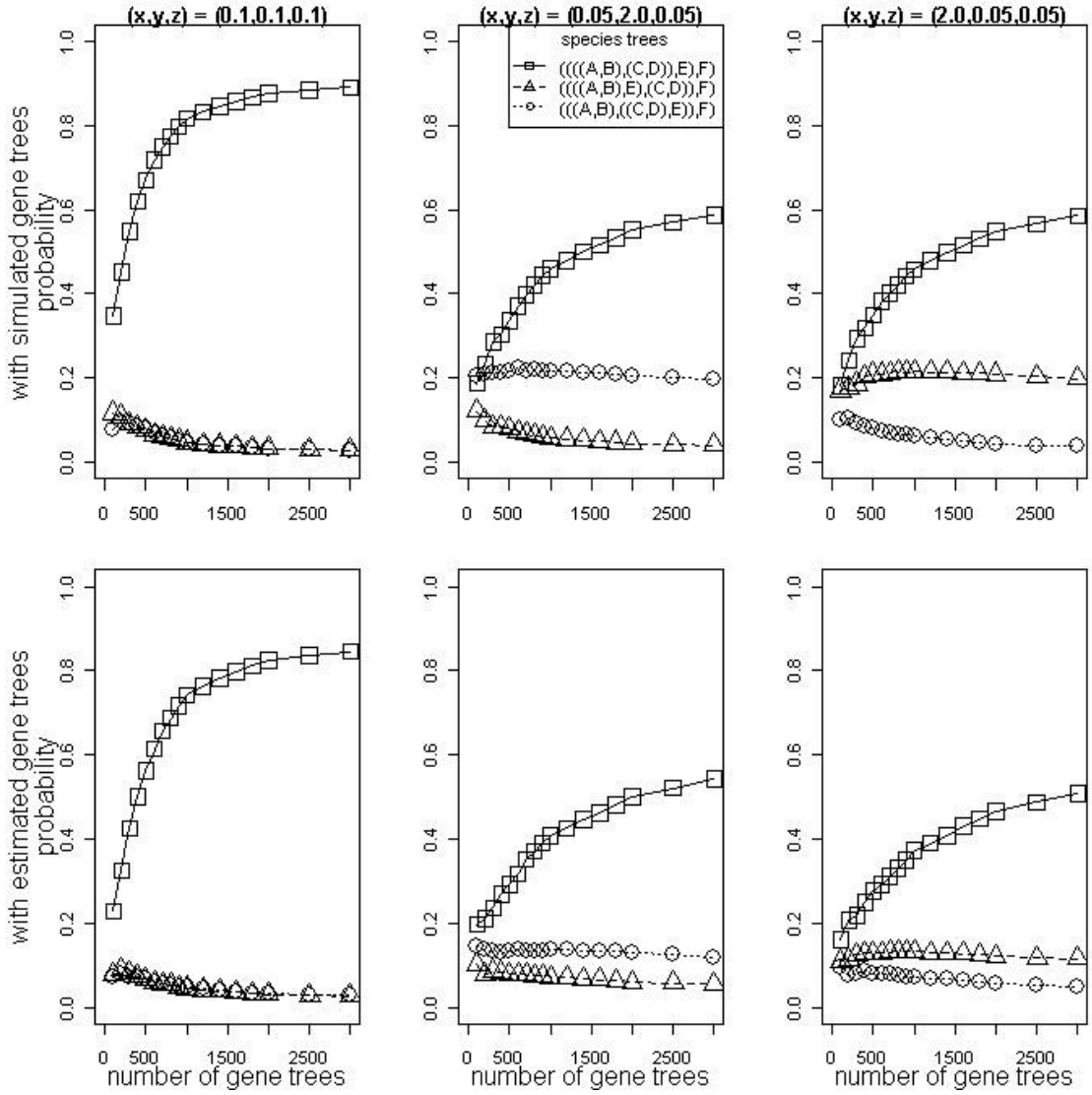


Figure 28. Pruning 1 taxon randomly under the true species tree in Figure 26. All plots have the same legend as the top middle one.

When pruning 1 taxon randomly from the simulated gene trees, the performance of MRP depended on the choice of branch lengths. For example, with 3000 simulated gene trees and pruning 1 taxon randomly instead of no pruning, the matching

estimated species tree topology ((((A,B),(C,D),E),F) was returned about 8% less often for the branch length (x, y, z) = (0.1, 0.1, 0.1); but about 20% more frequently for both (x, y, z) = (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05) (compare the top rows of Figures 27 and 28).

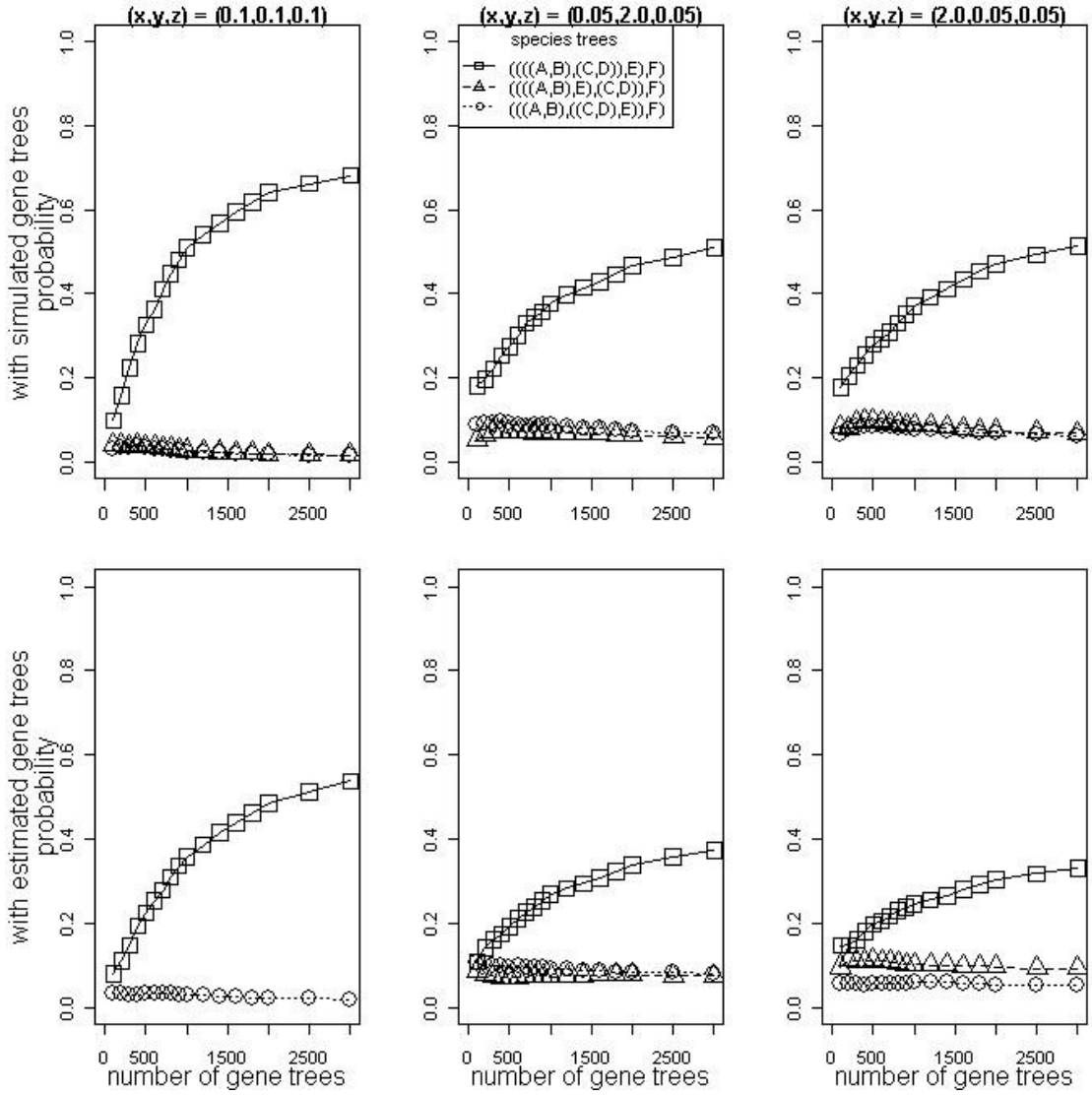


Figure 29. Pruning 2 taxa randomly under true the species tree in Figure 26. All plots have the same legend as the top middle one.

If 2 taxa are pruned from the gene trees, the matching estimated species tree topology was yielded less often but still more frequently than all the non-matching estimated species tree topologies, as demonstrated in Figure 29. For example, with 3000

simulated gene trees, the matching estimated species tree topology ((((A,B),(C,D),E),F) was returned about 21%, 8% and 7% less often with pruning 2 taxa than 1 taxon, for the branch lengths (x, y, z) = (0.1, 0.1, 0.1), (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05), respectively (compare the top rows of Figures 28 and 29). The same results can be found with the estimated gene trees under the same true species tree topology ((((A,B),(C,D),E),F) in the simulation (compare the bottom rows of Figures 28 and 29).

The effect of using estimated gene trees was dependent on the pruning schemes used under this true species tree topology ((((A,B),(C,D),E),F). On the one hand, with no pruning, the matching estimated species tree topology was returned more often (or about the same proportion of time), if using the estimated rather than simulated gene trees in all the selected branch lengths in the simulation (compare the top and bottom rows of Figure 27).

On the other hand, the most frequently outputted estimated species tree topology still matched the true species tree topology but less often, when at least 1 taxon was pruned (compare the top and bottom rows of Figures 28 and 29). For instance, the matching estimated species tree topology ((((A,B),(C,D),E),F) was returned about 15%, 13% and 18% less often, for the corresponding branch lengths (x, y, z) = (0.1, 0.1, 0.1), (0.05, 2.0, 0.05) and (2.0, 0.05, 0.05) with 3000 estimated rather than

simulated gene trees (compare top and bottom rows of Figure 29).

In conclusion, the simulation results suggested that under the true species tree model ((((A,B),(C,D),E),F) for selected branch lengths, (x, y, z) = (0.05, 2.0, 0.05) compared to (2.0, 0.05, 0.05), there was a relationship between the non-matching estimated species tree topologies (((A,B),((C,D),E)),F) and ((((A,B),E),(C,D)),F). For the selected branch lengths and when pruning at least 1 taxon, the matching estimated species tree topology ((((A,B),(C,D),E),F) was returned most often. However, with estimated gene trees and when no pruning is used, the matching estimated species tree topology was returned most often.

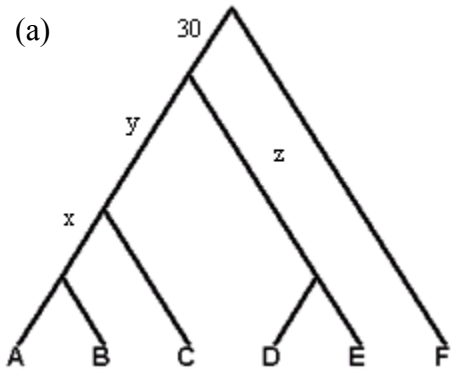
3.7.3 Under topology ((((A,B),C),(D,E)),F)

Figure 30 is the true species tree model with the branch lengths (x, y, z), and the possible MRP estimated species tree topologies. The following 2 sets of branch lengths measured in coalescent units (x, y, z) = (0.1, 0.1, 0.1) and (0.05, 0.05, 2.0) were used. Hence, there are 12 cases, 2 sets of branch lengths, 3 pruning schemes and 2 types of gene trees. Only 3 estimated species trees were reported to illustrate the performance of MRP.

All the most frequently outputted estimated species tree topologies contain both {AB} and {DE} clades. Also, from the simulation, the matching estimated species tree

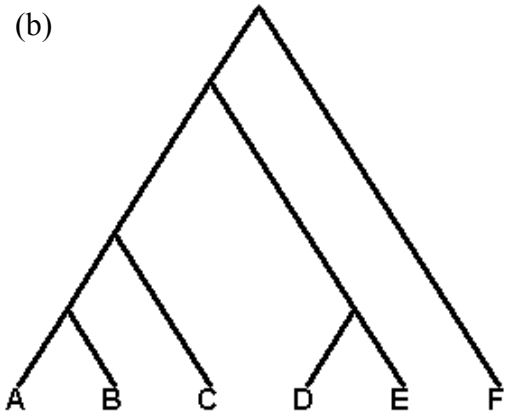
topology ((((A,B),C),(D,E)),F) was returned more often with uniformed spaced branch lengths (x, y, z) = (0.1, 0.1, 0.1) than (0.05, 0.05, 2.0). The non-matching matching estimated species topology tree was returned most often with the branch lengths (x, y, z) = (0.05, 0.05, 2.0), see top row of Figure 31.

(a)



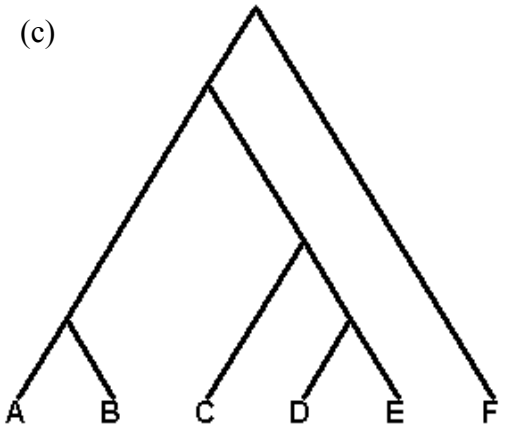
The species tree used in the simulation with branch lengths (x, y, z) and a taxon F as the outgroup with branch length of 30 coalescent units.

(b)



The matching MRP estimated species tree topology with the same outgroup.

(c)



A non-matching MRP estimated species tree topology with the same outgroup.

Figure 30. Species tree model and possible outcomes.

The performance of MRP under the species tree topology ((((A,B),C),(D,E)),F) depended on the different pruning schemes. For example, when using 3000 simulated gene trees, and the branch lengths (x, y, z) = (0.05, 0.05, 2.0), the matching estimated species tree topology was yielded about 30% more often when pruning 1 taxon

(compare the top right of Figures 31 and 32). However, the performance of MRP was different if pruning 2 taxa. For example, the matching estimated species tree topology was returned about 25% and 17% less frequently with 3000 simulated gene tree by having 2 taxa instead of 1 randomly pruned (compare the top rows of Figures 32 and 33). In either case, the matching estimated species tree topology ((((A,B),C),(D,E)),F) was returned most often when pruning at least 1 taxon from the simulated and estimated gene trees in the simulation.

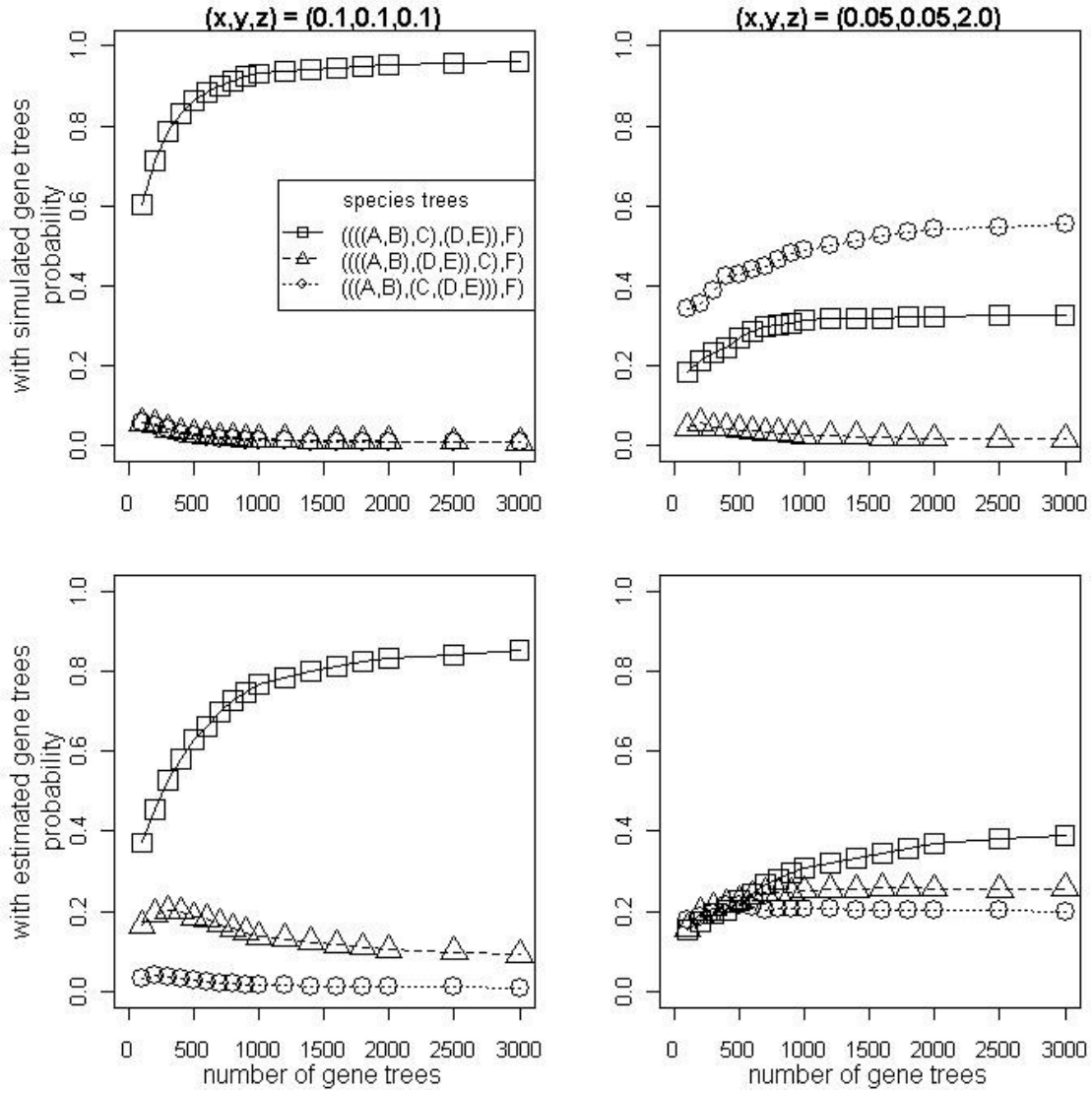


Figure 31. No pruning under the species tree in Figure 30. All plots have the same legend as the top one.

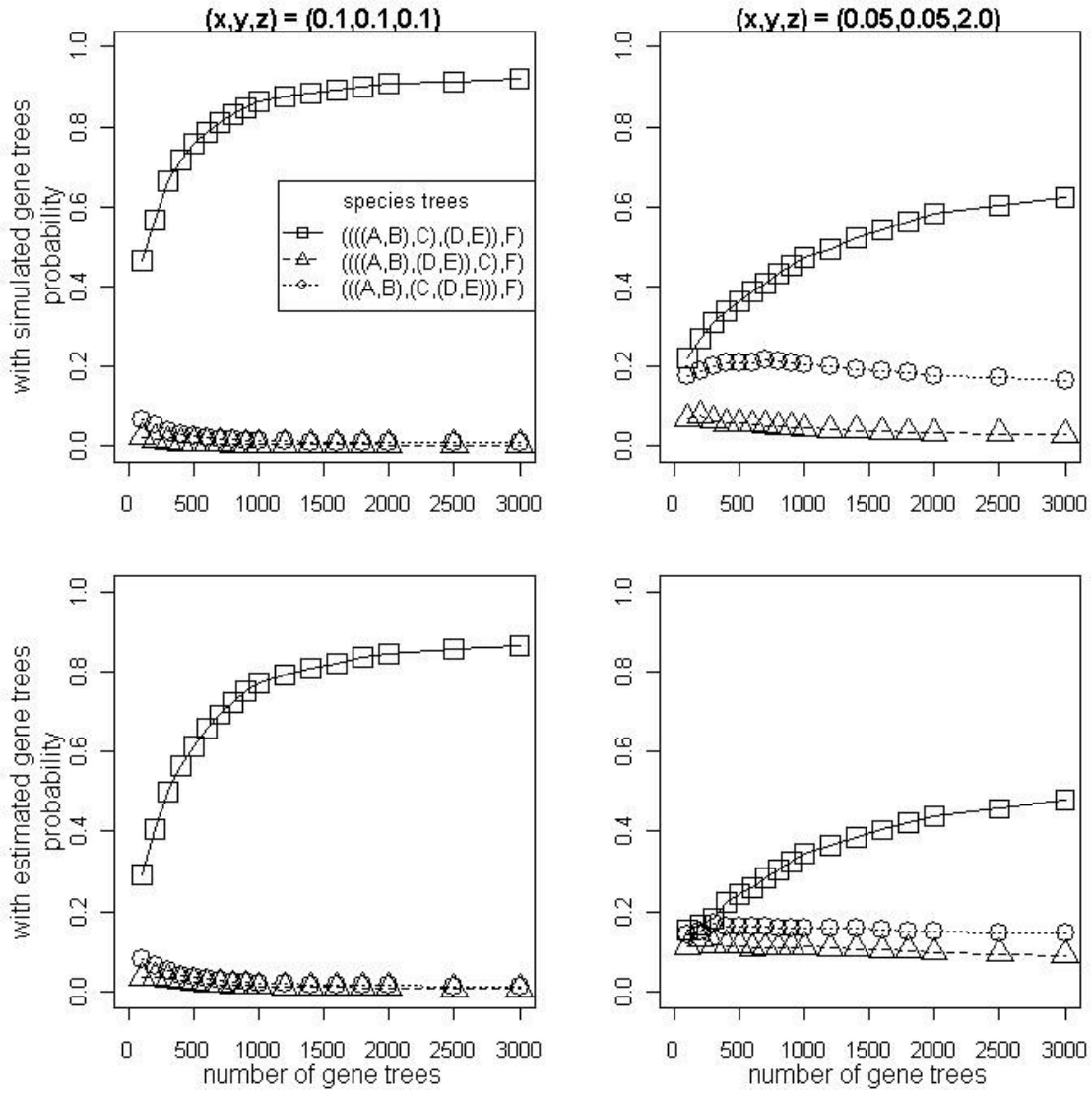


Figure 32. Pruning 1 taxon randomly pruned under the species tree in Figure 30. All plots have the same legend as the top one.

Similar to the effect of pruning schemes, the performance of MRP was different with the estimated versus simulated gene trees in various settings under the true species tree ((((A,B),C),(D,E)),F). For instance, with no pruning, the matching estimated species tree topology was returned about 10% less often for branch lengths (x, y, z) = (0.1, 0.1, 0.1), but about 6% more often for (x, y, z) = (0.05, 0.05, 2.0) when using 3000 estimated rather than simulated gene trees (compare the top and bottom rows of Figure 31). However, if pruning at least 1 taxon from the gene trees, the matching

estimated species tree topology was always yielded less frequently, when using the estimated instead of simulated gene trees, for the same branch lengths (compare the top and bottom rows of Figures 32 and 33).

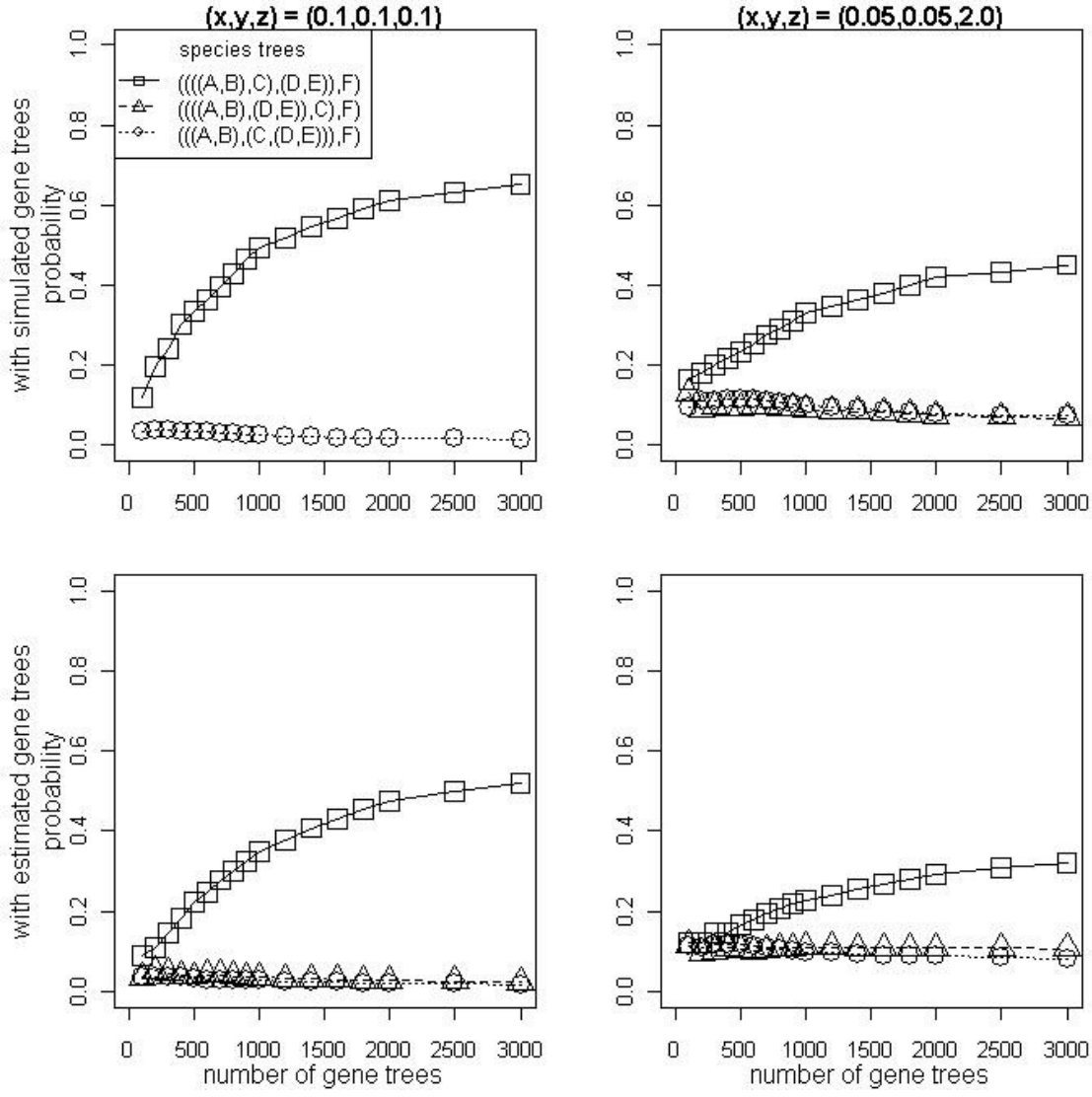


Figure 33. Pruning 2 taxa randomly under the species tree in Figure 30. All plots have the same legend as the top one.

There were some common simulation results from all the species tree topologies considered: (((((A,B),C),D),E),F), ((((A,B),(C,D),E),F) and ((((A,B),C),(D,E)),F). First, the matching estimated species tree topology was returned more often with uniformly distributed branch lengths, like (x, y, z) = (0.1, 0.1, 0.1). For the other

branch lengths, the matching estimated species tree topology was yielded more frequently, when pruning at least 1 taxon from the simulated and estimated gene trees. However, that the matching estimated species tree topology was returned less frequently, when more taxa are randomly pruned (i.e. pruning 2 taxa rather than 1) as less information is available.

The matching estimated species tree topology was returned more often using the estimated rather than simulated gene trees with no pruning, for some cases in the simulation. In contrast, for the branch lengths (x, y, z) = (0.1, 0.1, 0.1), the matching estimated species tree topology was returned more often for simulated than estimated gene trees in the simulation.

3.8 Species trees with 20 taxa and outgroup

The performance of MRP for true species tree topologies with 4 and 5 of taxa is demonstrated in previous sections (Sections 3.6 and 3.7). In this section, species trees with 20 taxa and an outgroup were used to explore the performance of MRP. There are more than 8.0×10^{21} bifurcating trees with 20 taxa (Felsenstein, 2004), and the branch lengths can be any positive value. Hence, it is hard to cover all the possible combinations.

Because of this, 20 species trees with 20 taxa (given in the Appendix) were randomly

generated with random branch lengths, using Mesquite (Maddison, 2009) under a Yule model (Yule, 1925) with a height of 10 coalescent units. An outgroup with 50 coalescent units was added to each simulated true species tree. The Yule model is an evolutionary model, and each species has an equal probability of undergoing a speciation event with a constant rate at any given point in time (Gernhard et al., 2008).

The simulation procedure is exactly the same as before (Section 3.1) but with 4 different pruning schemes: (i) no pruning; (ii) randomly deleting 25%; (iii) 50%; and (iv) 75% of the taxa. Here, a pruning scheme of randomly deleting 25% of taxa means that each taxon is pruned independently with probability of 0.25, so that on average, 25% of the taxa are pruned. However, it is possible to have more or less than 25% of taxa pruned from a single gene tree, and similarly for the cases of 50% and 75%. For each combination of true species tree and pruning scheme, the following numbers of gene trees were simulated: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2500 and 3000 with 300 replications in each case. Only some key results were shown to illustrate the performance of MRP and the rest can be found in the Appendix.

The normalized Robinson-Foulds distance was used to assess the performance of MRP. Recall that the maximal Robinson-Foulds distance of 2 bifurcating trees with

the same *n* taxa is $(n-2) \times 2$, as there are at most $n-1$ nodes and both trees have the same taxa set at the root. Hence, for 20-taxon trees, the corresponding maximal Robinson-Foulds distance is $(20-2) \times 2 = 36$, the calculated Robinson-Foulds distance at given iteration was normalized by dividing with 36 to give a value between 0 and 1. The error bar for each number of gene trees was also given to show the standard error of the normalized Robinson-Foulds distance in the simulation.

The MRP estimated species tree topology matches the true species tree topology only when the normalized Robinson-Foulds distance is exactly 0. However, this did not appear in any of the results (Figures 34 – 39). The reason for this observation is that even the normalized Robinson-Foulds distances between the estimated and the true species tree topologies was 0 for most of the time in the simulation, when there was at least 1 node difference between the topologies of the estimated and the true species tree in an iteration, the resulting normalized Robinson-Foulds distance was at least $(1+1)/36 = 1/18$. Because of this, the average of the normalized Robinson-Foulds distance across all iterations was close to 0 but never reached 0 in simulation.

Also, as the maximal Robinson-Foulds distance is 36 for the 20-taxon species tree, a normalized Robinson-Foulds distance of 0.1 in the plots indicated that, when comparing the topologies between the MRP estimated and the true species tree in the

simulation, there was on average about $36 \times 0.1/2 = 1.8$ nodes difference. This result implied that these 2 topologies matched for all but 2 nodes in the simulation, so that the performance of MRP was reasonable.

The results (Figures 34 – 39) shown that for each true species tree, the performance of MRP was about the same using simulated gene trees with no pruning and deleting 25% of the taxa, as the normalized Robinson-Foulds distance was roughly the same under both settings in the simulation. However, the performance of MRP was worse, if using simulated gene trees with pruning 50% and 75% of the taxa. For such cases, the normalized Robinson-Foulds distance was higher for small numbers of simulated gene trees initially. As the number of simulated gene trees increased, the normalized Robinson-Foulds distance decreased slowly.

MRP performed well for species trees 1 and 3 as the normalized Robinson-Foulds distance was below 0.05 given enough gene trees and decreased, as the number of simulated gene trees increased with all the pruning schemes (Figures 34 and 35). However, for trees 11 and 15 (Figures 37 and 38), with all the pruning schemes, the corresponding normalized Robinson-Foulds distance only decreased for small numbers of simulated gene trees. If using a larger number of simulated gene trees, the resulting normalized Robinson-Foulds distance remained approximately the same. Hence, for the true species tree like trees 11 and 15, the performance of MRP was not

improved necessarily by increasing the number of simulated gene trees.

However, for trees 7 and 17 (Figures 36 and 39), the normalized Robinson-Foulds distance (i.e. the performance of MRP) was roughly the same regardless of the number of simulated gene trees and the pruning schemes, and the performance of MRP is slightly worse if more taxa are pruned.

In conclusion, the results (Figures 34 – 39) shown that the performance of MRP can be improved in some cases by using a larger number of gene trees, so that the corresponding normalized Robinson-Foulds distance was close to 0 (the matching estimated species tree topology was returned more often). For example, with 3000 simulated gene trees and no pruning, the corresponding normalized Robinson-Foulds distance for trees 1 and 3 were 0.0167 and 0.0121, respectively from the simulation (Figures 34 and 35). Similarly, the performance of MRP was improved only with smaller numbers of simulated gene trees initially for some true species trees. If larger numbers of simulated gene trees was used, the performance was steady (e.g. trees 11 and 15 in Figures 37 and 38). In contrast, the normalized Robinson-Foulds distance was almost the same for different numbers of simulated gene trees under some true species trees (e.g. trees 7 and 17). To this end, the performance of MRP in such cases was stable in the simulation.

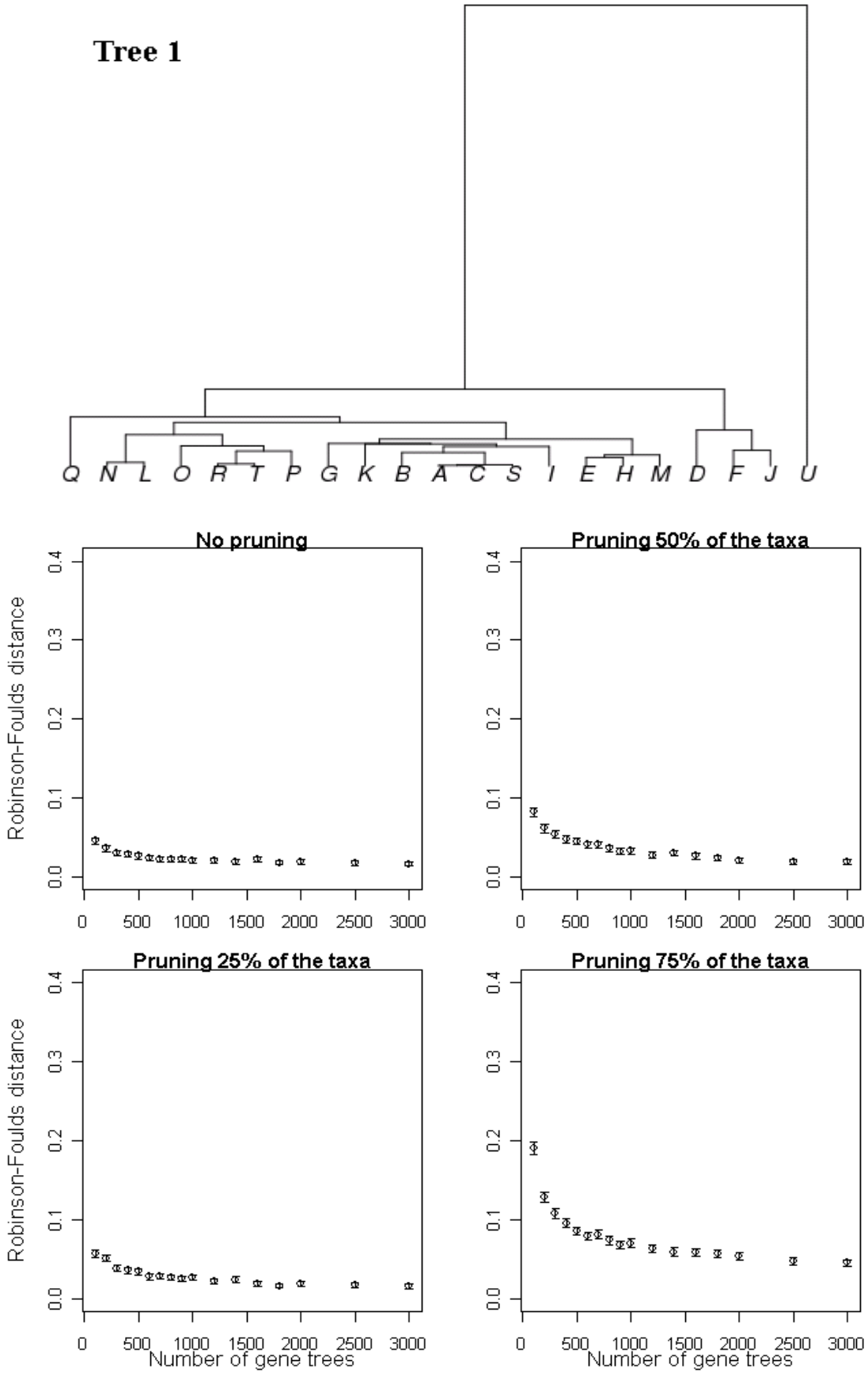


Figure 34. Tree 1 with simulated gene trees.

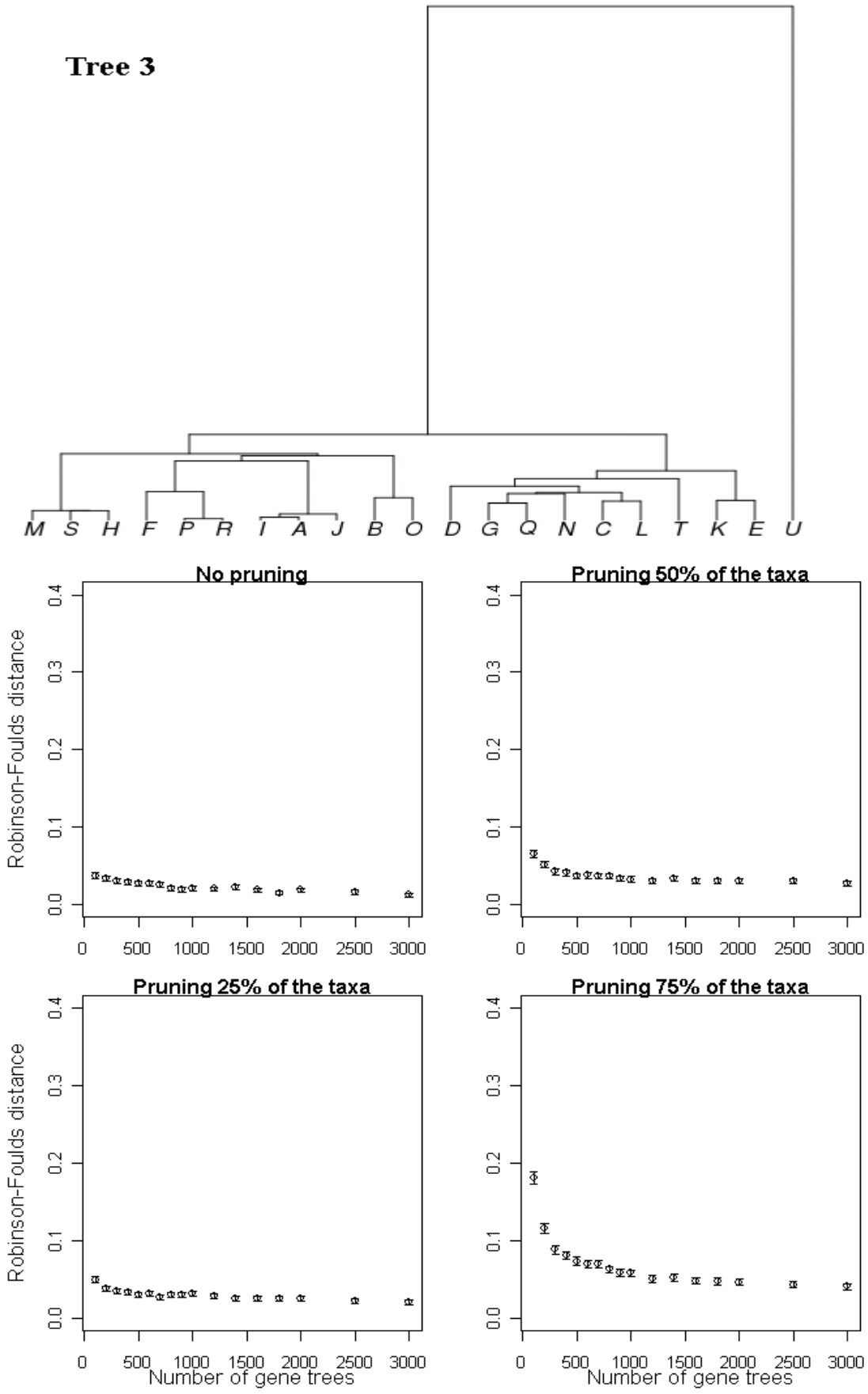


Figure 35. Tree 3 with simulated gene trees.

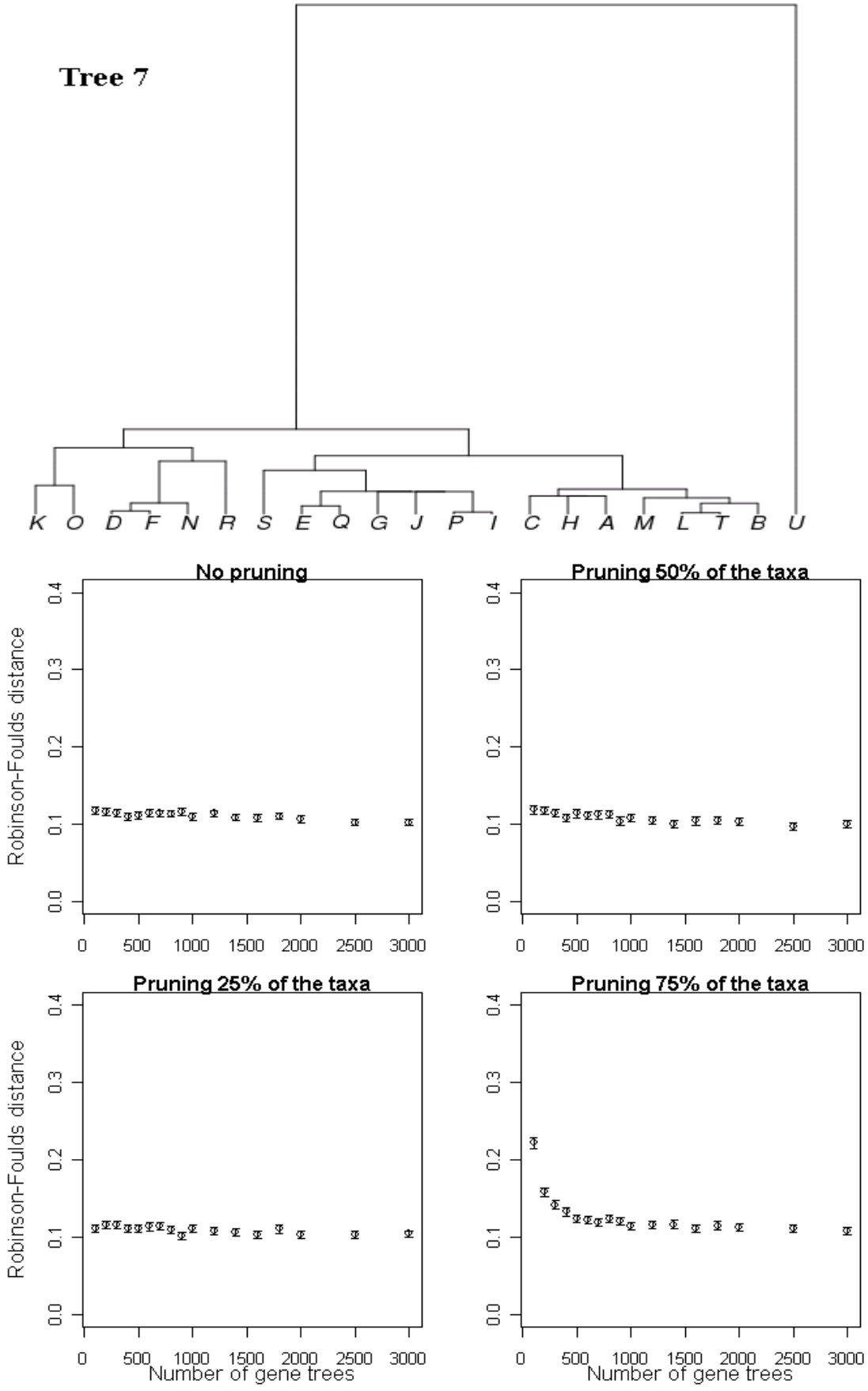


Figure 36. Tree 7 with simulated gene trees.

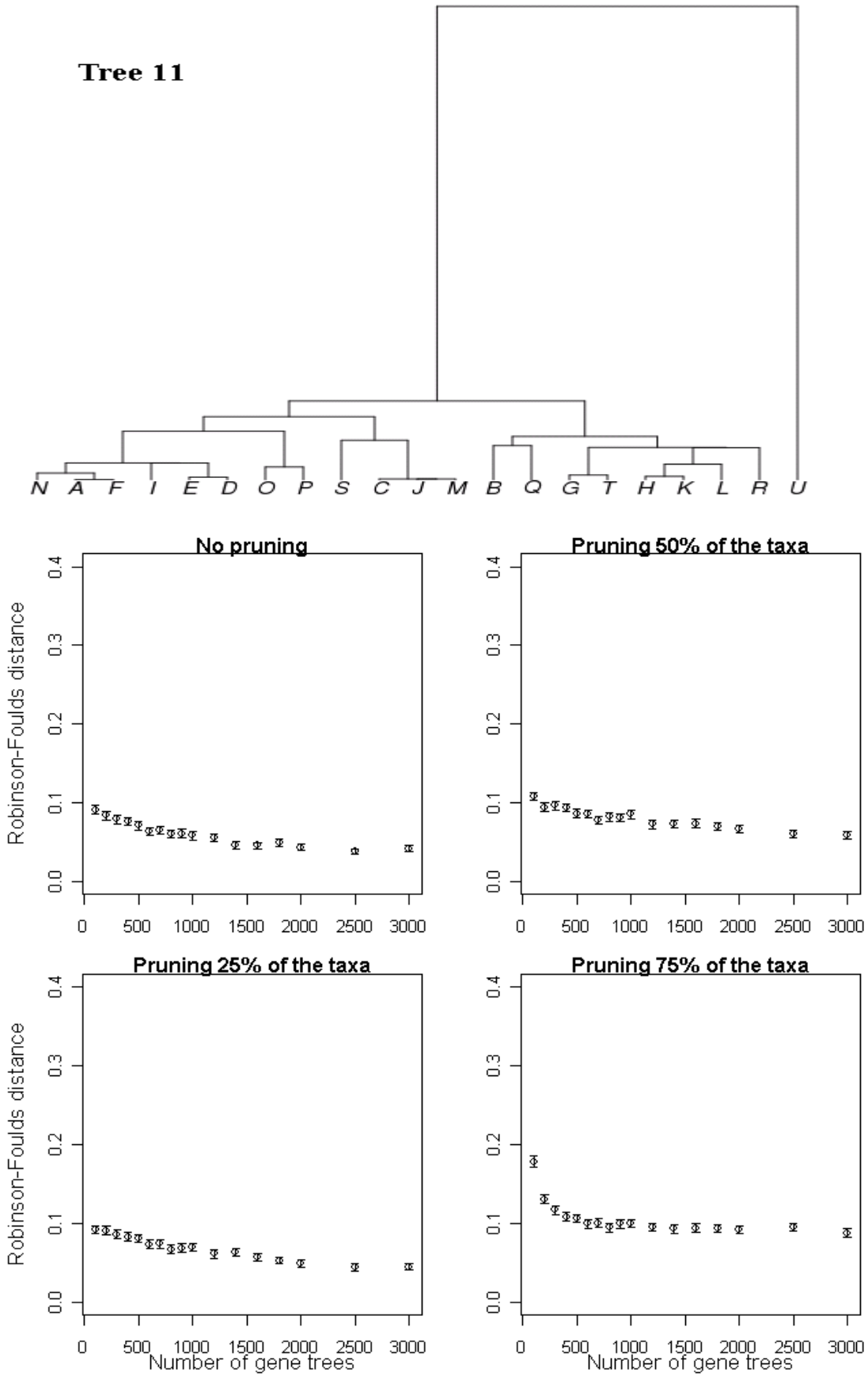


Figure 37. Tree 11 with simulated gene trees.

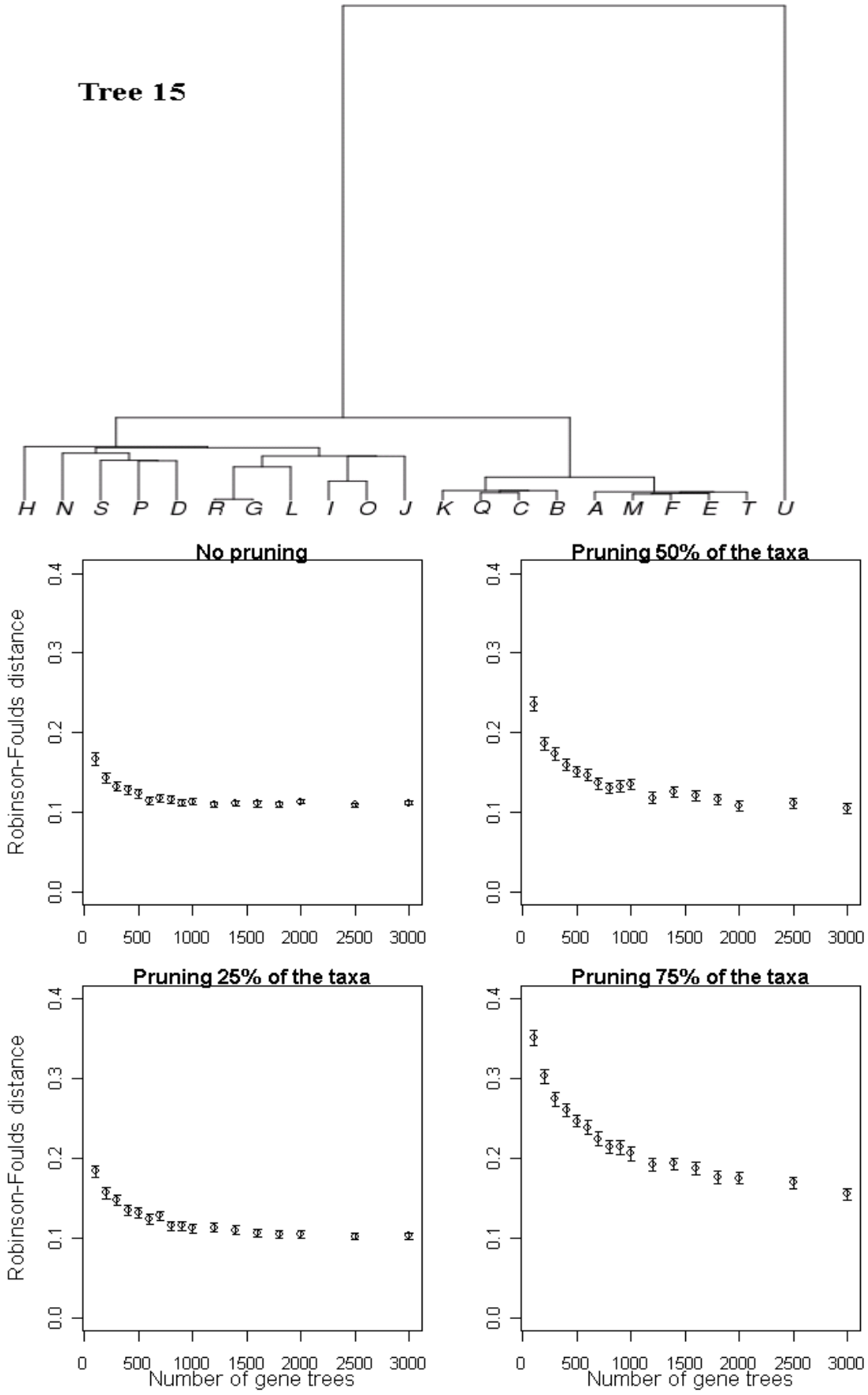


Figure 38. Tree 15 with simulated gene trees.

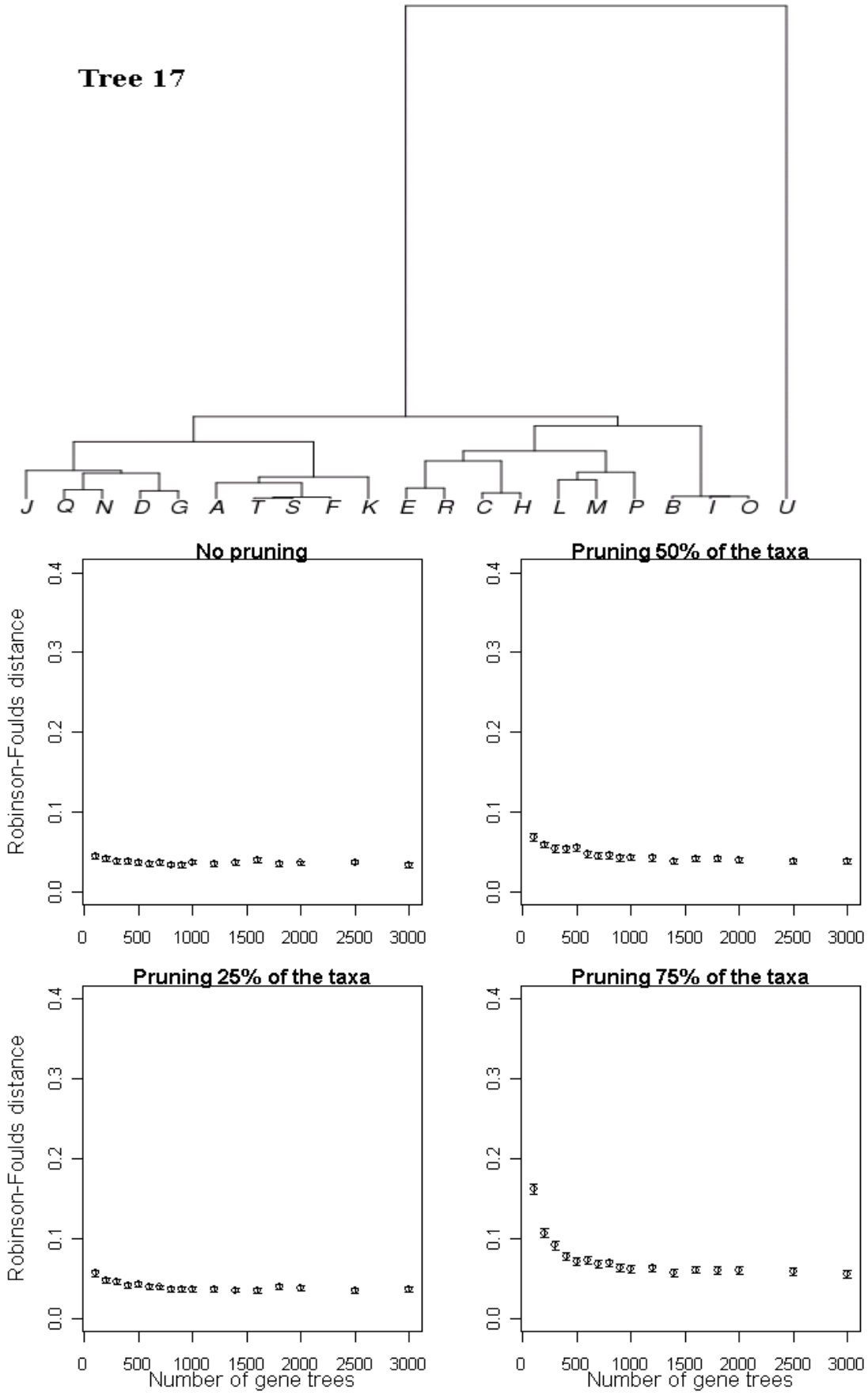


Figure 39. Tree 17 with simulated gene trees.

The performance of MRP was related to the true species trees, as shown in Sections 3.6 and 3.7 (i.e. 4-taxon and 5-taxon species trees). We investigated the effect of the 20-taxon species trees on the corresponding MRP performance (Figures 40 and 41). For each true 20-taxon species tree used in the simulation, the probability of corresponding matching gene tree was calculated (using COAL), and the shortest branch length was found (using R). The normalized Robinson-Foulds distance of 3000 simulated gene trees was plotted against results (Figures 40 and 41), using letters A, B, … , T in the alphabet order to represent the 20 species trees with 20 taxa used in the simulation. Thus, it is easy to see if there is a pattern that explains why the normalized Robinson-Foulds distance of some trees were higher than the others in the simulation.

Recall that there are more than 8.0×10^{21} bifurcating trees with 20 taxa. Consequently, a 1×10^{-4} probability of matching gene tree is relatively large. To this end, the corresponding probabilities of matching gene tree for trees 3, 11 and 17 (letters C, K and Q in Figure 40) are about 2×10^{-3}, 6×10^{-3} and 1.2×10^{-2}. Under these 3 true species trees in the simulation, the resulting normalized Robinson-Foulds distance was below 0.05. Equivalently, the matching estimated species tree topology under these 3 true species trees was returned more often (Figures 35, 37 and 39).

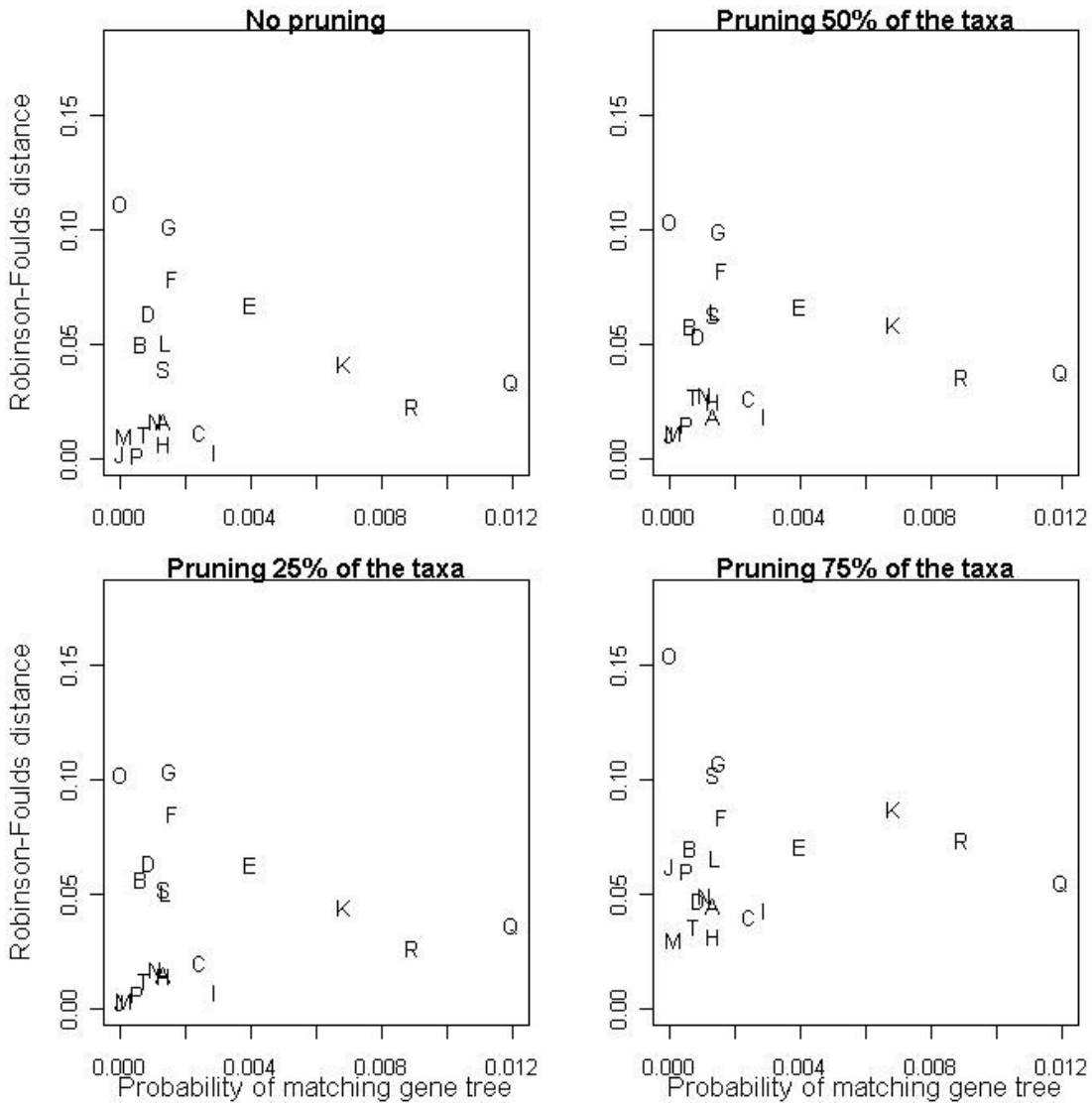


Figure 40. Probability of returning the matching gene tree for the 20-taxon species trees used.

Figure 41 illustrates that the shortest branch length of tree 1 (letter A) is 0.048, which is the longest among the 20-taxon species trees used in the simulation. In addition, its corresponding probability of matching gene tree was more than 0.001. Under the true species tree 1, the resulting normalized Robinson-Foulds distance was below 0.05 (Figure 34). However, there are relatively short branch lengths for both trees 7 and 15 (letters G and O in Figure 41). Moreover, the corresponding probabilities of matching gene trees were also low (Figure 40). For these trees, the

normalized Robinson-Foulds distance was bigger than 0.1, as shown in the simulation (Figures 36 and 38), so that the matching estimated species tree topology was returned relatively less often.

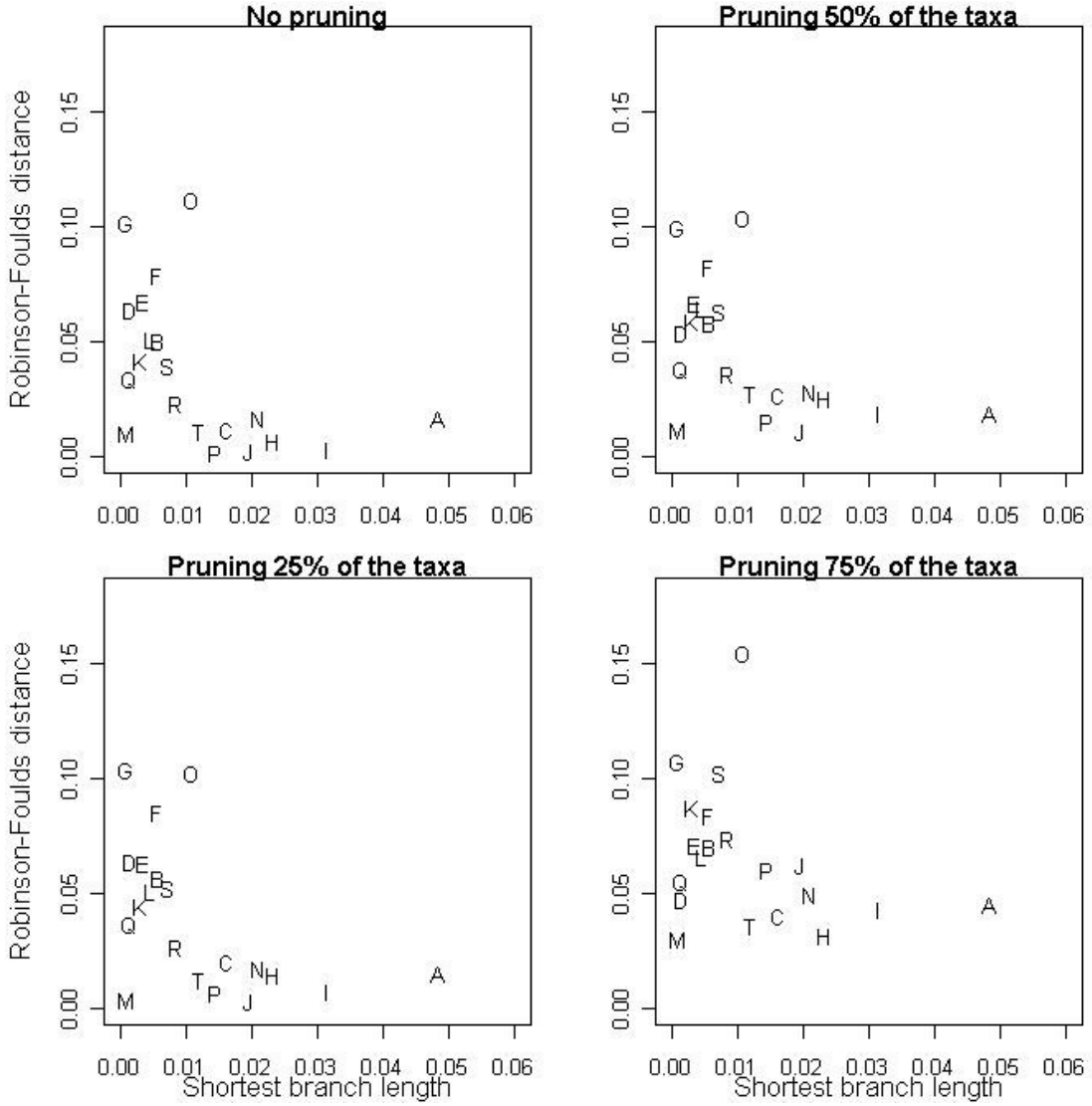


Figure 41. Shortest branch length of the 20-taxon species trees used.

On the one hand, Figure 41 shows that both trees 17 and 7 (letters Q and G) including a relatively short branch length. On the other hand, Figure 40 indicates that the probability of returning the matching gene tree for tree 17 is higher than for tree 7. Although the simulation results indicated a steady performance under the true species trees 7 and 17, the corresponding normalized Robinson-Foulds distance of tree 17 was

lower than tree 7 under the same setting. In other words, the matching estimated species tree topology under tree 17 was returned more often than tree 7 under the same simulation setting. A similar observation can be found for trees 11 and 15 (letters K and O).

In contrast, there is a low normalized Robinson-Foulds distance (less than 0.05) for tree 13 (letter M), and with a relatively short branch length and low probability of returning the matching gene tree (Figures 40 and 41). This observation indicates that other tests can be used to assess the effect of the 20-taxon species trees on the corresponding MRP performance.

Next, the performance of MRP was explored further with estimated gene trees. The estimated gene trees were obtained using the same mutation model as for the 4 and 5 taxa species trees with outgroup (Section 3.3). The only pruning scheme used was no pruning, as it was computationally expensive to convert 20-taxon trees into DNA sequences, and then estimate gene trees from the resulting DNA sequences. Again, the same numbers of gene trees were used: 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2500 and 3000 with 300 replications. The normalized Robinson-Foulds distance was used to assess the performance of MRP in the simulation.

Figure 42 shows that the performance of MRP was worse with estimated gene trees compared to simulated gene trees, i.e. the normalized Robinson-Foulds distance was higher for all the true species trees used.

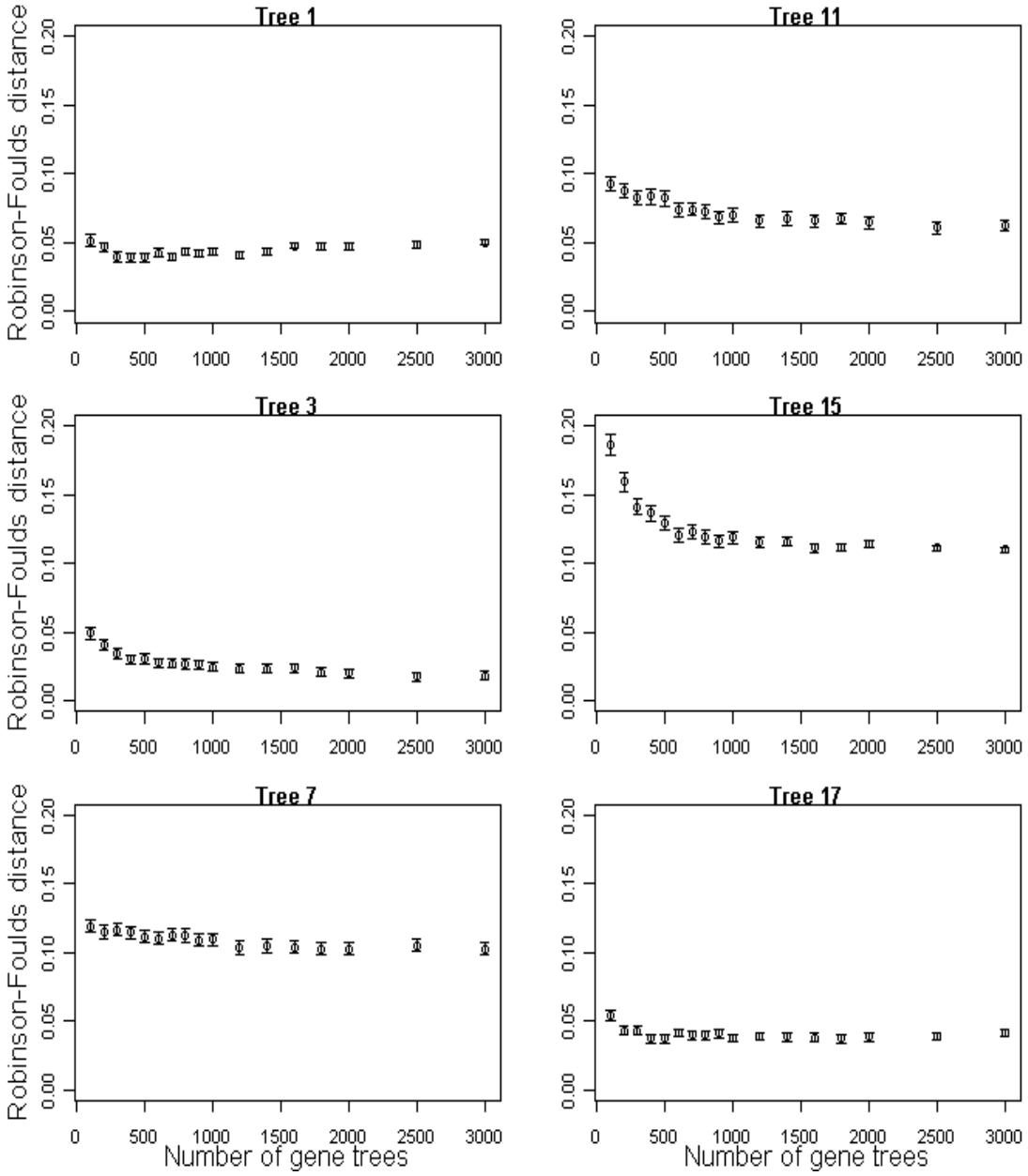


Figure 42. With estimated gene trees and no pruning.

In summary, in the simulation, the performance of MRP was related to the 20-taxon species trees. On the one hand, the performance of MRP for true species trees 1 and 3 can be improved by using more simulated gene trees (resulting in a lower normalized

Robinson-Foulds distance). On the other hand, for true species trees 11 and 15, the performance of MRP was not improved by using more simulated gene trees. In addition, when pruning a large set of taxa from the simulated gene trees, there was less information available. Consequently, the matching estimated species tree topology was returned less often so that the corresponding normalized Robinson-Foulds distance was further far away from 0. When using estimated rather than simulated gene trees in the simulation, the performance of MRP was worse, as the corresponding normalized Robinson-Founds distance was higher.

4 Analytical study

Chapter 3 explores the performance of MRP using simulation. In this chapter, we employ analytical equations to calculate the expected parsimony scores of candidate trees, to study the performance MRP from a different perspective.

MRP returns the estimated species tree topology with the lowest total parsimony score summed across all input gene trees, so that as the number of gene trees increases towards infinity, the corresponding total parsimony score sum also increases towards infinity. However, the average parsimony score converges to the expected value of the parsimony score (EPS) asymptotically (by the law of large number), and computing the EPS can help us to determine the MRP asymptotic supertree (MRPAST) under the particular setting of interest. Then by comparing MRPAST with the true species tree topology, one can easily tell if a matching estimated species tree topology can be returned by MRP under certain conditions in the limit as the number of gene trees approaches infinity.

4.1 Implementation

One can think of the gene trees as having a discrete distribution with 2 types of parameters, species tree topology and branch lengths. The topology depicts the relationships of n taxa, and there are $n - 2$ internal branch lengths to consider. The

same gene tree under the same species tree topology could have different probabilities of occurring purely because of different branch lengths.

Mutation is not considered here because it is difficult to obtain the distribution of gene trees that are estimated from DNA sequences. Pruning schemes are allowed such that all the gene trees are informative as described in Section 3.2. The pruning schemes can yield 2 settings: consensus and supertree. In the consensus setting, no taxa are pruned, and in the supertree setting, at least 1 taxon is pruned. Let *w* denote the probability of deleting at least 1 taxon from the gene tree so that the weight of the consensus setting is 1 – *w* and the weight of supertree setting is *w*, and *w* is a value between 0 and 1. *w* = ½ indicates that each gene tree is equally likely to have 0 taxa pruned or at least 1 randomly pruned.

We let EPS denote the expected parsimony score, which is a function of (i) the true species tree topology of interest with *n* taxa, *topo*; (ii) the branch lengths of the topology, $b_1, b_2, \cdots, b_{n-2}$; (iii) the pruning scheme weight factor *w* and (iv) the pruning scheme with the probability of deleting each possible taxon, $d_1, d_2, \cdots, d_n$ which sum up to 1. For *m* possible gene trees under the species tree topology, we denote the corresponding probabilities of occurring each gene tree as $p_1, p_2, \cdots, p_m$. These probabilities depend on the topology and branch lengths of the true species tree. In the supertree setting, $p_1, p_2, \cdots, p_m$ also depend on the probabilities of deleting

taxa $d_1, d_2, \cdots, d_n$ (Degnan and Salter, 2005). Then, the form of EPS is:

$EPS\left(topo, b_1, b_2, \cdots, b_{n-2}, w, d_1, d_2, \cdots, d_n\right) = (1-w) \times EPS_C + w \times EPS_S$, where

$EPS_C = \left(c_1 p_1 + c_2 p_2 + c_3 p_3 + \cdots + c_m p_m\right)$ and

$EPS_S = \left(s_1 p_1 + s_2 p_2 + s_3 p_3 + \cdots + s_m p_m\right)$.

The terms c_1, s_1, c_2, s_2 up to c_m, s_m are the parsimony scores in the consensus and supertree settings of all *m* corresponding possible gene trees under the true species tree topology of interest. That is, each gene tree probability for a given topology is weighted with its corresponding parsimony score within the pruning scheme and then the EPS is the sum of all these terms that are weighted according to the pruning scheme weight factor *w* (An example is given in the Appendix).

It is easy to calculate the *n*th moment of EPS as

$EPS_C^{(n)} = \left(c_1^n p_1 + c_2^n p_2 + c_3^n p_3 + \cdots + c_m^n p_m\right)$ and

$EPS_S^{(n)} = \left(s_1^n p_1 + s_2^n p_2 + s_3^n p_3 + \cdots + s_m^n p_m\right)$ so that the *n*th moment of EPS is

$EPS^{(n)}\left(topo, b_1, b_2, \cdots, b_{n-2}, w, d_1, d_2, \cdots, d_n\right) = (1-w) \times EPS_C^{(n)} + w \times EPS_S^{(n)}$.

To use the EPS equation to find the MRPAST, a true species tree topology is selected. Then, for each candidate tree *t*, the corresponding EPS_t is computed by setting the same branch lengths and pruning scheme of interest. This leads to the candidate tree (or more than 1 if there is a tie) with the lowest EPS_t being the MRPAST for the given

selection of the true species tree, branch lengths and pruning scheme. This result can be used to check if the MRPAST matches the true species tree topology or not.

It is quite hard to compute the EPS for topologies with many taxa. In this analytical study, only topologies of 4 and 5 taxa are considered, for which there are 15 and 105 possible bifurcating candidate trees, respectively (Felsenstein, 2004). The same 2 topologies (((A,B),C),D) and ((A,B),(C,D)) are covered for the 4-taxon case as in last chapter. Similarly, ((((A,B),C),D),E), (((A,B),(C,D)),E) and (((A,B),C),(D,E)) are the only 3 topologies considered for the 5-taxon case. All the EPS equations of each case are keyed and run in MAPLE (Maplesoft, Inc.)

4.2 Species trees with 4 taxa

To be consistent with the Chapter 3, the same 4-taxon species tree models (excluding an outgroup) are used (Figures 16 and 19).

Under the true species tree topology (((A,B),C),D), the non-matching estimated species tree topology, ((A,B),(C,D)) can be returned most frequently, as shown in the simulation (Figure 17). Therefore, the task is to find when the estimated species tree matches the true species tree topology (((A,B),C),D) in the limit as the number of gene trees approaches infinity. That is, the task is to find when the EPS_t for candidate tree (((A,B),C),D) is strictly less than for candidate tree ((A,B),(C,D)) with the same

combination of the branch lengths (x, y) and the pruning schemes.

Let (((A,B),C),D) be the true species tree topology and denote the candidate trees (((A,B),C),D) and ((A,B),(C,D)) as *t1* and *t13* (the full list can be found in the Appendix). Let EPS_{t1} and EPS_{t13} be the EPS_t for *t1* and *t13*. To ensure the gene tree is informative, at most 1 taxon can be pruned. Thus, the EPS_{t1} and EPS_{t13} are calculated for the all combinations of the branch lengths (x, y), the pruning scheme weight factor and the pruning schemes by using MATLAB (Mathwork, Inc.).

The 6 curves in Figure 43 represents the 6 pruning schemes using 6 different pruning scheme weight factors w: 0, 0.2, 0.4, 0.6, 0.8 or 1 so that the first curve from the top is when $w = 0$ (i.e. the consensus setting) and so forth. For any of the 6 curves, $EPS_{t1} = EPS_{t13}$ (i.e. a tie) when using any branch lengths (x, y) on the curves. Similarly, when using any branch lengths (x, y) below a particular curve, the MRPAST is the non-matching species tree topology *t13* = ((A,B),(C,D)), i.e. $EPS_{t1} > EPS_{t13}$ under current pruning scheme. For instance, for any branch lengths (x, y) below the first curve, the corresponding MRPAST is *t13* = ((A,B),(C,D)) with infinitely many simulated gene trees in the consensus setting (i.e. $w = 0$). In contrast, when using any branch lengths (x, y) above particular curve, the MRPAST is the matching species tree topology *t1* = (((A,B),C),D), i.e. $EPS_{t1} < EPS_{t13}$ with the given pruning scheme.

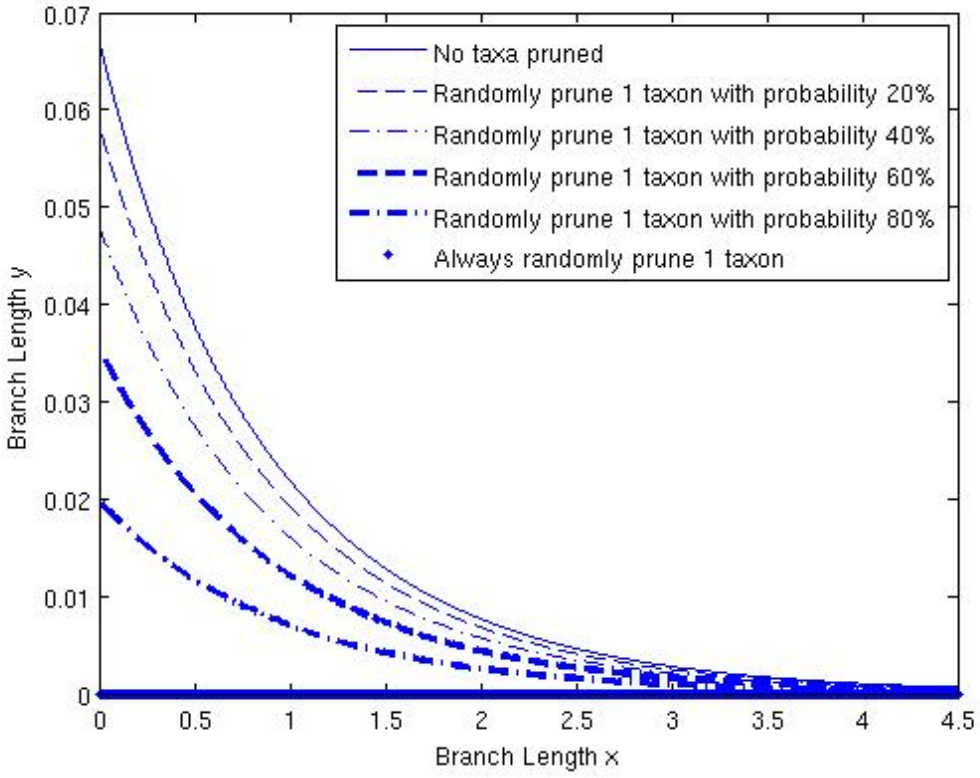


Figure 43. The true species tree topology is (((A,B),C),D).
The ranges of the branch lengths (x, y) are 0.001 to 0.07 and 0.001 to 4.5
with a step size of 2×10^{-5} in coalescent units
(excluding 0 as the branch lengths must be positive to form a bifurcating species tree).
It is equally likely (or randomly) to prune any of the taxa when necessarily.

Figure 43 also illustrates that when the number of simulated gene trees approaches infinity, the MRPAST is $t1$ = (((A,B),C),D) for more combinations of branch lengths (x, y) by using a larger proportion of simulated gene trees with 1 taxon pruned. This is the same as increasing the pruning scheme weight factor w towards 1, so that the corresponding area below the curve is reduced. Consequently, there exists more possible combinations of branch lengths (x, y) such that $EPS_{t1} < EPS_{t13}$. To this end, under the true species tree topology (((A,B),C),D), when always pruning 1 taxon from infinitely many simulated gene trees (w = 1), the resulting MRPAST matches the true species topology always (because the bottom curve of Figure 43 is on the horizontal

axis). The simulation result (Figure 18) agrees with this observation.

The symbolic equation of the curves in Figure 43 can be solved for x, by setting $EPS_{t1} = EPS_{t13}$:

$$EPS_{t1} = (1-w)\left(2+\frac{2}{3}e^{-x}+\frac{2}{3}e^{-y}+\frac{1}{3}e^{-x-y}-\frac{1}{18}e^{-x-3y}\right) + w\left[\left(1+\frac{2}{3}e^{-y}\right)(d_A+d_B)+\left(1+\frac{2}{3}e^{-x-y}\right)d_C+\left(1+\frac{2}{3}e^{-x}\right)d_D)\right]; \quad (4.1)$$

$$EPS_{t13} = (1-w)\left(3+\frac{2}{3}e^{-x}-\frac{1}{3}e^{-y}+\frac{1}{6}e^{-x-y}+\frac{1}{18}e^{-x-3y}\right) + w\left[\left(2-\frac{1}{3}e^{-y}\right)(d_A+d_B)+\left(1+\frac{2}{3}e^{-x-y}\right)d_C+\left(1+\frac{2}{3}e^{-x}\right)d_D)\right]; \quad (4.2)$$

so that

$$x = \log\left[\frac{(3e^{2y}-2)(1-w)}{18(e^{3y}-e^{2y})\left(1-w(1-d_A-d_B)\right)}\right]. \quad (4.3)$$

Here, w is the pruning scheme weight factor that indicates the probability of pruning taxon from the gene trees; d_A and d_B are the probabilities of deleting taxa A and B, respectively; and x, y are the branch lengths. If w is 0, 0.2, 0.4, 0.6, 0.8 or 1; and $d_A = d_B = d_C = d_D =$¼ , is the same as randomly pruning 1 taxon. Then, by solving x in terms of y for equation (4.3), one can precisely re-generate the curves in Figure 43.

Therefore, under the true species topology ((((A,B),C),D), for the selections of w, d_A and d_B , with infinitely many simulated gene trees, the MRPAST does not match the true species topology when for a given branch length y, the branch length x is strictly less than the x-value calculated from equation (4.3) so that $EPS_{t1} > EPS_{t13}$. Thus, for any combinations of branch lengths (x, y) satisfying inequality (4.4), the resulting

MRPAST does not match the true species tree topology ((((A,B),C),D).

$$x < \log\left[\frac{(3e^{2y}-2)(1-w)}{18(e^{3y}-e^{2y})\big(1-w(1-d_A-d_B)\big)}\right]. \quad (4.4)$$

From the inequality (4.4), when $w = 1$, (always pruning 1 taxon from the input gene tree, supertree setting), then the numerator of inequality (4.4) is 0. Consequently, the right hand side of inequality (4.4) becomes log(0) such that there is no solution. Hence, there is no combination of (x, y) satisfying inequality (4.4). Equivalently, MRPAST matches the true species tree topology (((A,B),C),D) for all the combinations of (x, y) in the supertree setting with infinitely many simulated gene trees. This gives an explanation of the counterintuitive result discovered in the simulation (Figure 18). That is, the matching estimated species tree topology (((A,B),C),D) is returned most frequently for any branch lengths, when pruning 1 taxon from all the infinitely many simulated gene trees.

Notice that inequality (4.4) only depends on d_A and d_B but not d_C or d_D. Hence, under the true species tree topology (((A,B),C),D), when using simulated gene trees with 1 taxon pruned ($w > 0$), the resulting MRPAST matches the true species tree topology more likely if deleting taxon A or B rather than taxon C or D. This is because that when $w > 0$ and $(1 - d_A - d_B)$ is large, for the same branch length y, it is more likely to have a branch lengths x larger than the corresponding x-value calculated from inequality (4.4). In other words, there are less combinations of branch length (x, y) satisfying inequality (4.4), so that deleting taxon A or B from the simulated gene trees

makes it is more likely to return a MRPAST that matches the true species tree topology (((A,B),C),D).

When $w = 0$, or $w \neq 0$ but the probabilities of deleting taxa A and B are both 0, $d_A = d_B = 0$, (no pruning taxa, consensus setting) then expression (4.4) becomes:

$$x < \log\left[\frac{(3e^{2y} - 2)}{18(e^{3y} - e^{2y})}\right]. \tag{4.5}$$

If the branch lengths satisfy inequality (4.5) when the true species tree topology is (((A,B),C),D), then this condition for the MRPAST to be ((A,B),(C,D)) with infinitely many simulated gene trees is identical to the condition that the asymptotic greedy consensus returns the non-matching estimated species tree topology ((A,B),(C,D)) rather than the matching estimated species tree topology (Degnan et al., 2009). This is equivalent to the area under the first curve from the top of Figure 43.

Because $(1 - d_A - d_B) \leq 1$, and $0 \leq w \leq 1$,

it follows that $\left(1 - w(1 - d_A - d_B)\right) \geq (1 - w)$

so that $\frac{(1-w)}{(1-w(1-d_A-d_B))} \leq 1$. Because logarithms are monotonic, inequality (4.6) follows:

$$\begin{aligned} x &< \log\left[\frac{(3e^{2y} - 2)(1 - w)}{18(e^{3y} - e^{2y})\left(1 - w(1 - d_A - d_B)\right)}\right] \\ &\leq \log\left[\frac{(3e^{2y} - 2)}{18(e^{3y} - e^{2y})}\right]. \end{aligned} \tag{4.6}$$

Expression (4.6) shows a relationship between greedy consensus and MRP for the true species tree (((A,B),C),D). That is, with infinitely many simulated gene trees, the same tree is returned from both methods with the same branch lengths (x, y) in the consensus setting. The matching topology is returned more easily by MRP than by the greedy consensus method in the supertree setting under the true species tree topology (((A,B),C),D).

For the rest of this section, we use the EPS to explore why the matching species tree topology is always returned under the true species tree topology ((A,B),(C,D)) for all the branch lengths and the pruning schemes. Hence, the task is to show that the EPS_{t13} of the candidate tree *t13* = ((A,B),(C,D)) is always the lowest compared to the other candidate trees (see the full ordered list in Appendix) under the same settings.

Again, w is the pruning scheme weight factor, and d_A, d_B, d_C, and d_D are the probabilities of deleting taxa A, B, C and D from infinitely many simulated gene trees. Let $X = e^{-x}$ and $Y = e^{-y}$ to simplify the EPS equations. Then, the EPS_t of all the candidate trees under the true species tree topology ((A,B),(C,D)) are given below, which is a collection of 9 groups:

$$EPS_{t1} = (1-w)(3 + 5/18\,XY - 1/3\,Y + 2/3\,X) + w[(2 - 1/3\,Y)(d_A + d_B) + (1 + 2/3\,X)(d_C + d_D)]; \quad (4.7)$$

$$EPS_{t2} = (1-w)(3 + 5/18\,XY - 1/3\,Y + 2/3\,X) + w[(2 - 1/3\,Y + 1/9\,XY)(d_A + d_B) + (1 + 2/3\,X)(d_C + d_D)]; \quad (4.8)$$

$$EPS_{t3} = EPS_{t7} = (1-w)(4 - 1/18\,XY - 1/3\,Y) + w[(2 - 1/3\,Y)(d_A + d_B) + (1 + 2/3\,X)d_C + (2 - 1/3\,X)d_D]; \quad (4.9)$$

$$EPS_{t4} = EPS_{t6} = (1-w)(4 - 1/18\,XY - 1/3\,X) + w[(1 + 2/3\,Y)d_A + (2 - 1/3\,y)d_B + (2 - 1/3\,X)(d_C + d_D)]; \quad (4.10)$$

$$EPS_{t5} = EPS_{t9} = (1-w)(4 - 1/18\,XY - 1/3\,Y) + w[(2 - 1/3\,Y)(d_A + d_B) + (2 - 1/3\,X)d_C + (1 + 2/3\,X)d_D]; \quad (4.11)$$

$$EPS_{t8} = EPS_{t10} = (1-w)(4 - 1/18\,XY - 1/3\,X) + w[(2 - 1/3\,Y)d_A + (1 + 2/3\,Y)d_B + (2 - 1/3\,X)(d_C + d_D)]; \quad (4.12)$$

$$EPS_{t11} = EPS_{t12} = (1-w)(3 + 5/18\,XY + 2/3\,Y + 1/3\,X) + w[(1 + 2/3\,Y)(d_A + d_B) + (2 - 1/3\,X)(d_C + d_D)]; \quad (4.13)$$

$$EPS_{t13} = (1-w)(2 + 2/9\,XY + 2/3\,Y + 2/3\,X) + w[(1 + 2/3\,Y)(d_A + d_B) + (1 + 2/3\,X)(d_C + d_D)]; \quad (4.14)$$

$$EPS_{t14} = EPS_{t15} = (1-w)(4 - 4/9\,XY) + w[(2 - 1/3\,Y)(d_A + d_B) + (2 - 1/3\,X)(d_C + d_D)]; \quad (4.15)$$

Here, the aim is to show that the EPS_{t13} of *t13*, i.e. equation (4.14), is the smallest compared to the other 8 groups so that the corresponding MRPAST matches the true species tree topology ((A,B),(C,D)). That is, one needs to show that the difference between equation (4.14) and any of the other equations is always negative under the same combinations of the branch lengths (x, y), the probabilities of deleting each possible taxon d_A, d_B, d_C, and d_D and the pruning scheme weight factor w.

The following is the proof.

Equation (4.14) – equation (4.7):

$$-(1-w)(1 + 1/18\ XY - Y) - w(d_A + d_B)(1 - Y). \quad (4.16)$$

Equation (4.14) – equation (4.8):

$$-(1-w)(1 + 1/18\ XY - Y) - w(d_A + d_B)(1 + 1/9\,XY - Y). \quad (4.17)$$

Equation (4.14) – equations (4.9):

$$\begin{aligned} &-(1-w)(2-5/18\ XY-2/3\,X-Y) \\ &-w[(d_A+d_B)(1-Y)+d_D(1-X)]. \end{aligned} \tag{4.18}$$

Equation (4.14) – equations (4.10):

$$\begin{aligned} &-(1-w)(2-5/18\ XY-X-2/3\,Y) \\ &-w[d_B(1-Y)+(d_C+d_D)(1-X)]. \end{aligned} \tag{4.19}$$

Equation (4.14) – equations (4.11):

$$\begin{aligned} &-(1-w)(2-5/18\ XY-2/3\,X-Y) \\ &-w[(d_A+d_B)(1-Y)+d_C(1-X)]. \end{aligned} \tag{4.20}$$

Equation (4.14) – equations (4.12):

$$\begin{aligned} &-(1-w)(2-5/18\ XY-X-2/3\,Y) \\ &-w[d_A(1-Y)+(d_C+d_D)(1-X)]. \end{aligned} \tag{4.21}$$

Equation (4.14) – equations (4.13):

$$-(1-w)(1+1/18\ XY-X)-w(d_C+d_D)(1-X). \tag{4.22}$$

Equation (4.14) – equations (4.15):

$$\begin{aligned} &-(1-w)(2-2/3\ XY-2/3\ X-2/3\ Y) \\ &-w[(d_A+d_B)(1-Y)+(d_C+d_D)(1-X)]. \end{aligned} \tag{4.23}$$

The expressions (4.16) to (4.23) can be shown to be negative because the branch lengths (x, y) are non-negative; and $X = e^{-x}$ and $Y = e^{-y}$, $0 < X, Y < 1$ so that $1 - X > 0$, $1 - Y > 0$ and $0 < XY < 1$. In addition, the terms d_A, d_B, d_C, and d_D are non-negative and sum to 1, and $0 \leq w \leq 1$. For instance, in equation (4.23),

$(d_A + d_B)(1 - Y) + (d_C + d_D)(1 - X) > 0$, and

$(2 - 2/3\ XY - 2/3\ X - 2/3\ Y) > (2 - 2/3 - 2/3 - 2/3) = 0.$

As $0 \leq w \leq 1$, expression (4.23) is always negative, and similarly this is true for expressions (4.16) to (4.22).

Therefore, all the expressions (4.16) to (4.23) are always negative under the true species topology ((A,B),(C,D)) for any combination of the branch lengths (x, y); the pruning scheme with d_A, d_B, d_C, and d_D as the probabilities of deleting taxon A, B, C and D and the pruning scheme weight factor w. Hence, the EPS_t of *t13* is always the lowest. As a result, the MRPAST is always *t13* = ((A,B),(C,D)), for which matches the true species topology ((A,B),(C,D)) as the number of simulated gene trees approaches infinity.

The asymptotic greedy consensus returns an estimated species tree that matches the true species tree topology ((A,B),(C,D)) for any combination of branch length (x, y) (Degnan et al., 2009). Hence, in the consensus setting and under the true species tree ((A,B),(C,D)), MRP and the greedy consensus method is equivalent, as the same tree is returned by both methods that matches the true species tree topology ((A,B),(C,D)) with infinitely many simulated gene trees.

4.3 Species trees with 5 taxa

For the 5-taxon species trees, 3 true species topologies are considered: ((((A,B),C),D),E); (((A,B),(C,D)),E) and (((A,B),C),(D,E)), which are the same species tree models but without the outgroup as in Chapter 3 (Figures 22, 26 and 30).

For the true species tree topology ((((A,B),C),D),E), the most frequently returned

estimated species trees from the simulation results in Section 3.7, *t1* = ((((A,B),C),D),E), *t61* = (((A,B),(C,D)),E) and *t76* = (((A,B),C),(D,E)), are considered. Similarly, under the true species tree topology (((A,B),(C,D)),E), the EPS_t are calculated for the most often estimated species tree returned in the corresponding simulation results, i.e. *t61* = (((A,B),(C,D)),E), *t88* = (((A,B),E),(C,D)) and *t103* = ((A,B),(E,(C,D))). In the same way, if the true species tree topology is (((A,B),C),(D,E)), the matching topology *t76* = (((A,B),C),(D,E)) and the most frequently returned non-matching topology *t105* = ((A,B),(C,(D,E))) from the simulation are used as the candidates to calculate the EPS_t.

Because there are 3 branch lengths (x, y, z), to generate the 2-D plot like Figure 43, z is fixed and then x and y are varied. Here, the pruning schemes are: (i) no pruning, (ii) always randomly pruning 1 taxon and (iii) pruning 2 taxa. Notice that these 3 pruning schemes are the same used for the 5-taxon topologies in the simulation (Section 3.7). Also, we assume that the probability of deleting each taxon A, B, C, D and E is equal so that $d_A = d_B = d_C = d_D = d_E = \frac{1}{5}$. The results are coloured to indicate the MRPAST for different settings. More results with other pruning schemes and the full ordered 105 candidate species trees list can be found in the Appendix.

In the plots (Figures 44 – 56), the white area is where the resulting MRPAST matches the corresponding true species topology. The non-white area is where the MRPAST

does not match. For instance, under the true topology ((((A,B),C),D),E), using no pruning, the MRPAST, *t1* = ((((A,B),C),D),E) matches the true species tree topology when using a branch lengths within the white area, e.g. (x, y, z) = (2, 0.1, 0.06). In contrast, if using a point in the non-white area, e.g. (x, y, z) = (2, 0.1, 0.05), the resulting MRPAST is the non-matching topology, *t76* = (((A,B),C),(D,E)), as shown in the Figure 44(b). All the simulation results in Chapter 3 for 5-taxon species tree agreed with the corresponding MRPAST results.

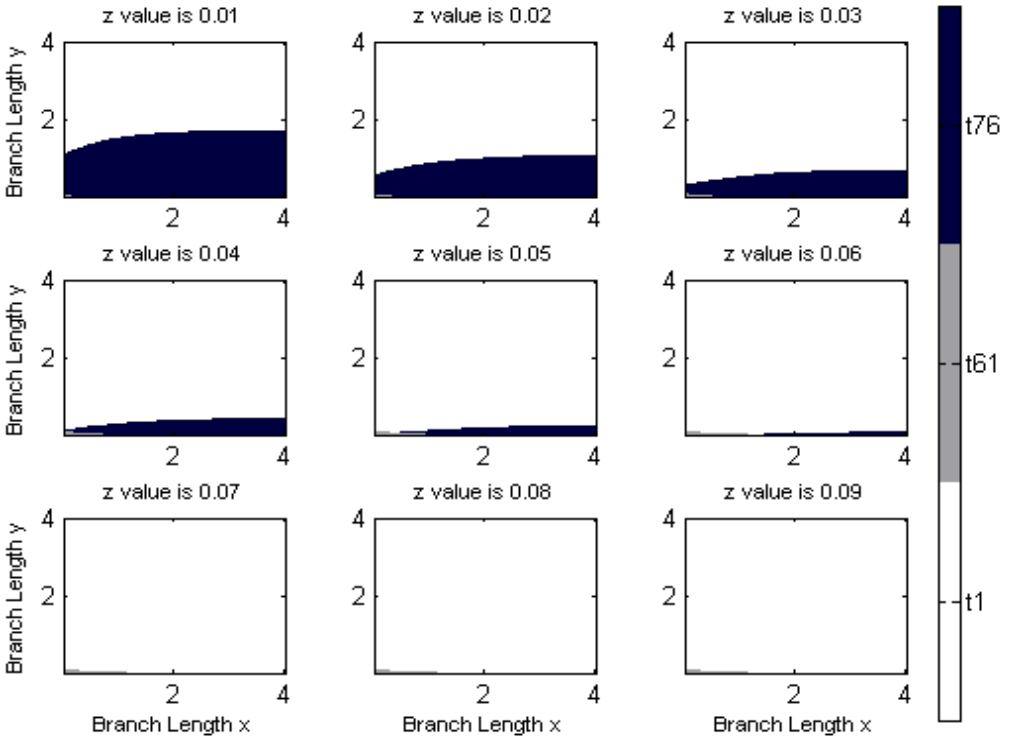
Under all these 3 true topologies: ((((A,B),C),D),E); (((A,B),(C,D)),E) and (((A,B),C),(D,E)), the MRPAST matches the true species tree topology more easily when pruning at least 1 taxon, i.e. the white area in the corresponding plots is larger. In addition, if pruning 2 taxa from the gene tree to form a rooted triple, the resulting MRPAST always the matches the true species tree topology regardless of the branch lengths so that the entire plots are white (Figures 48, 51 and 56).

Under the true topology ((((A,B),C),D),E), the matching topology is returned more easily with a longer branch lengths of y and z under the same pruning scheme (Figure 44). Similarly, for the true topology (((A,B),(C,D)),E), using a longer branch length of z, more combinations of x and y result in the MRPAST being the matching topology, *t61* = (((A,B),(C,D)),E) (Figure 49). However, under the true topology (((A,B),C),(D,E)), it is less easy to yield a MRPAST that matches the true species tree

topology with a longer branch length of *z* (Figure 53).

Another observation is that the relationship between the branch lengths *x* and *y* under the topology (((A,B),(C,D)),E), providing that *z* is small can change the MRPAST. On the one hand, a non-matching topology *t88* = (((A,B),E),(C,D)) is returned, if using a combination of longer *x* and shorter *y*. On the other hand, if using a combination of shorter *x* and longer *y*, a non-matching topology *t103* = ((A,B),(E,(C,D))) is returned (compare the dark and grey area in Figures 49 and 50).

(a)



(b)

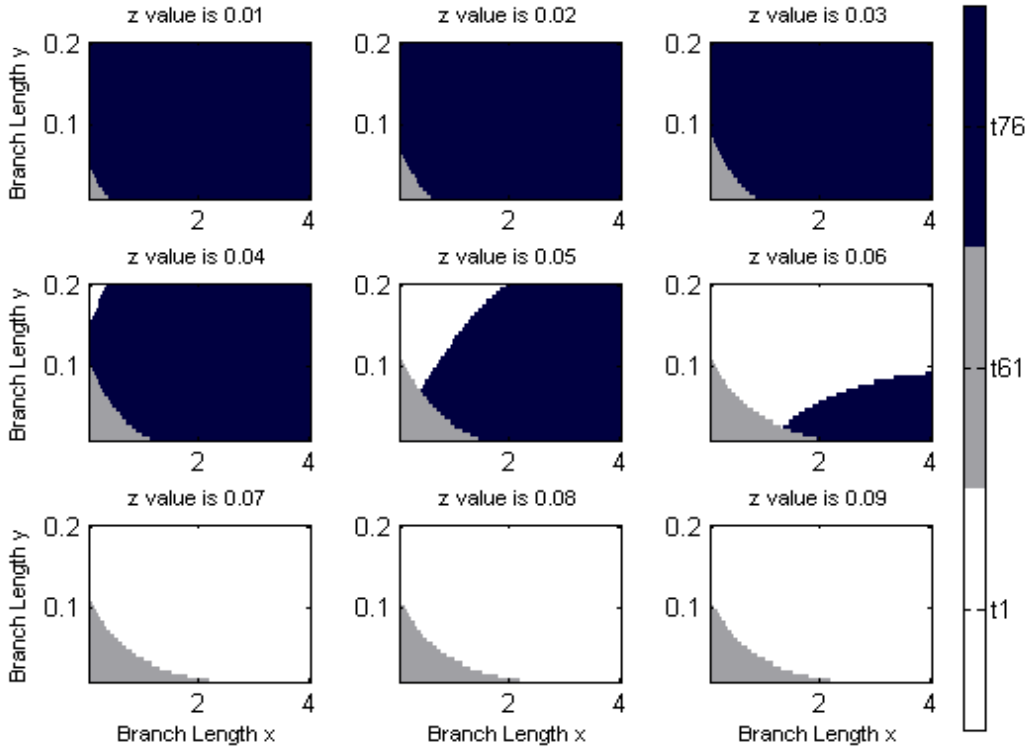


Figure 44. No pruning
under the true species tree topology t1=((((A,B),C),D),E).
Plot (b) is a zoom in of plot (a).

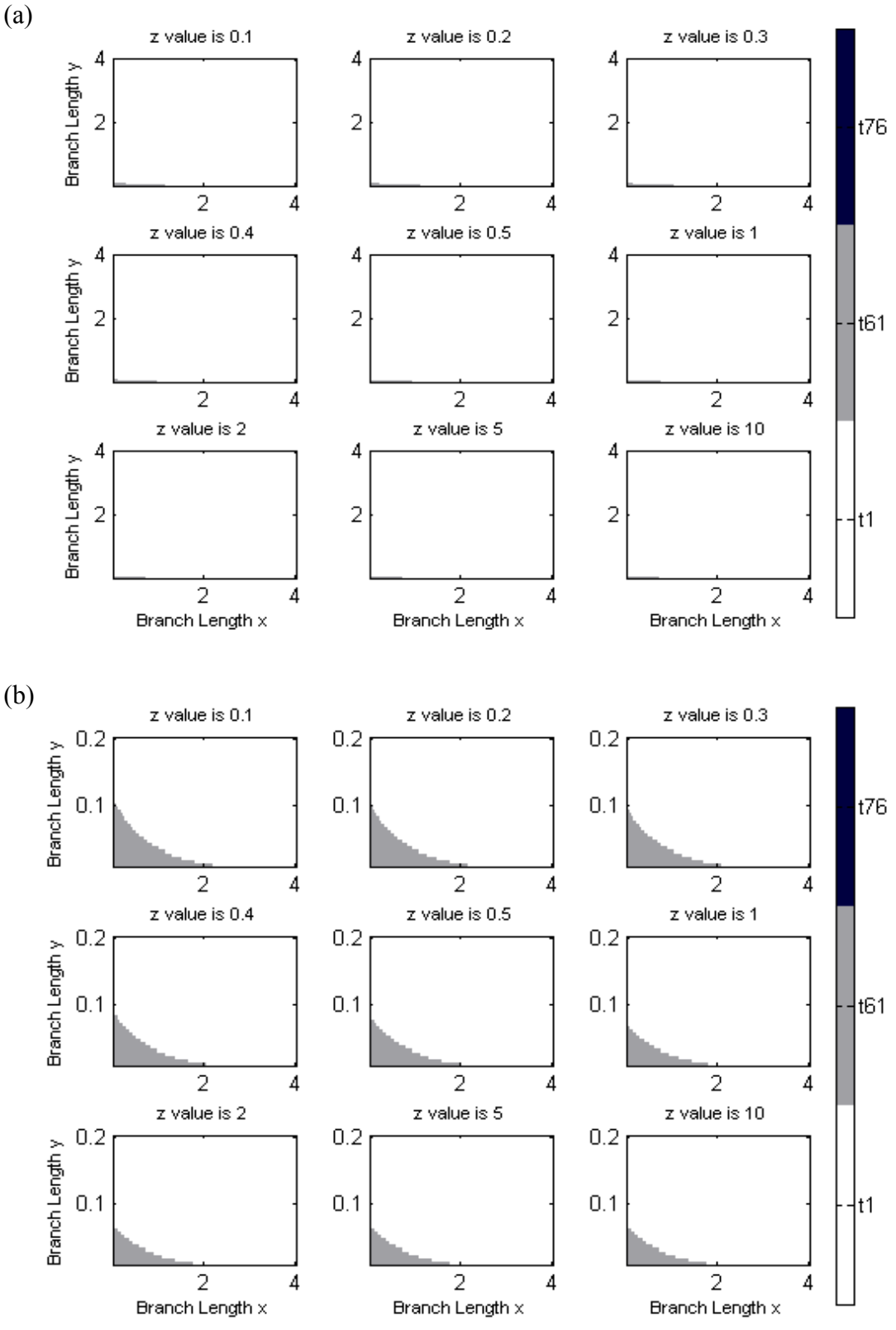


Figure 45. No pruning
under the true species tree topology t1=((((A,B),C),D),E).
Plot (b) is a zoom in of plot (a).

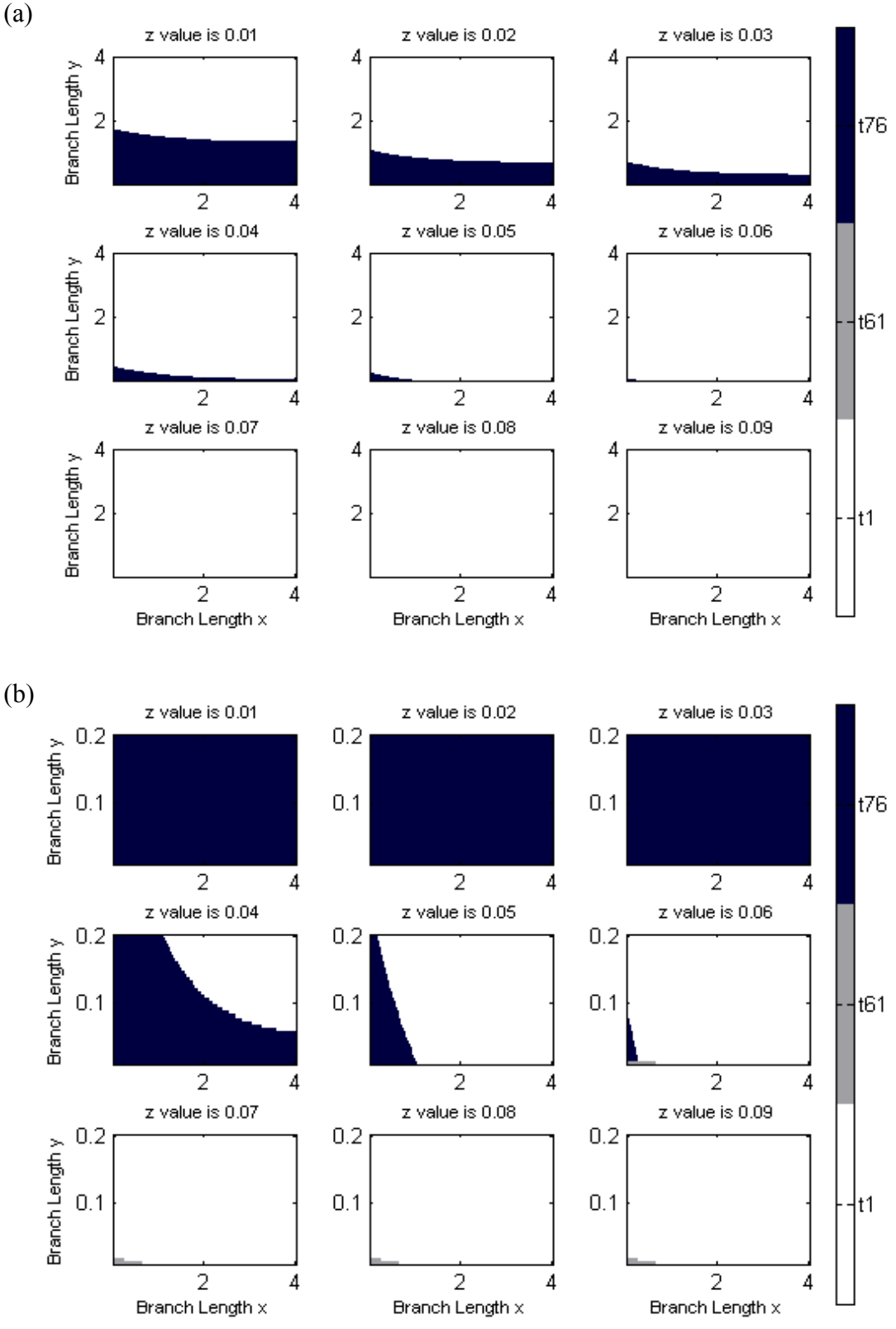


Figure 46. Pruning 1 taxon randomly
under the true species tree topology t1=((((A,B),C),D),E).
Plot (b) is a zoom in of plot (a).

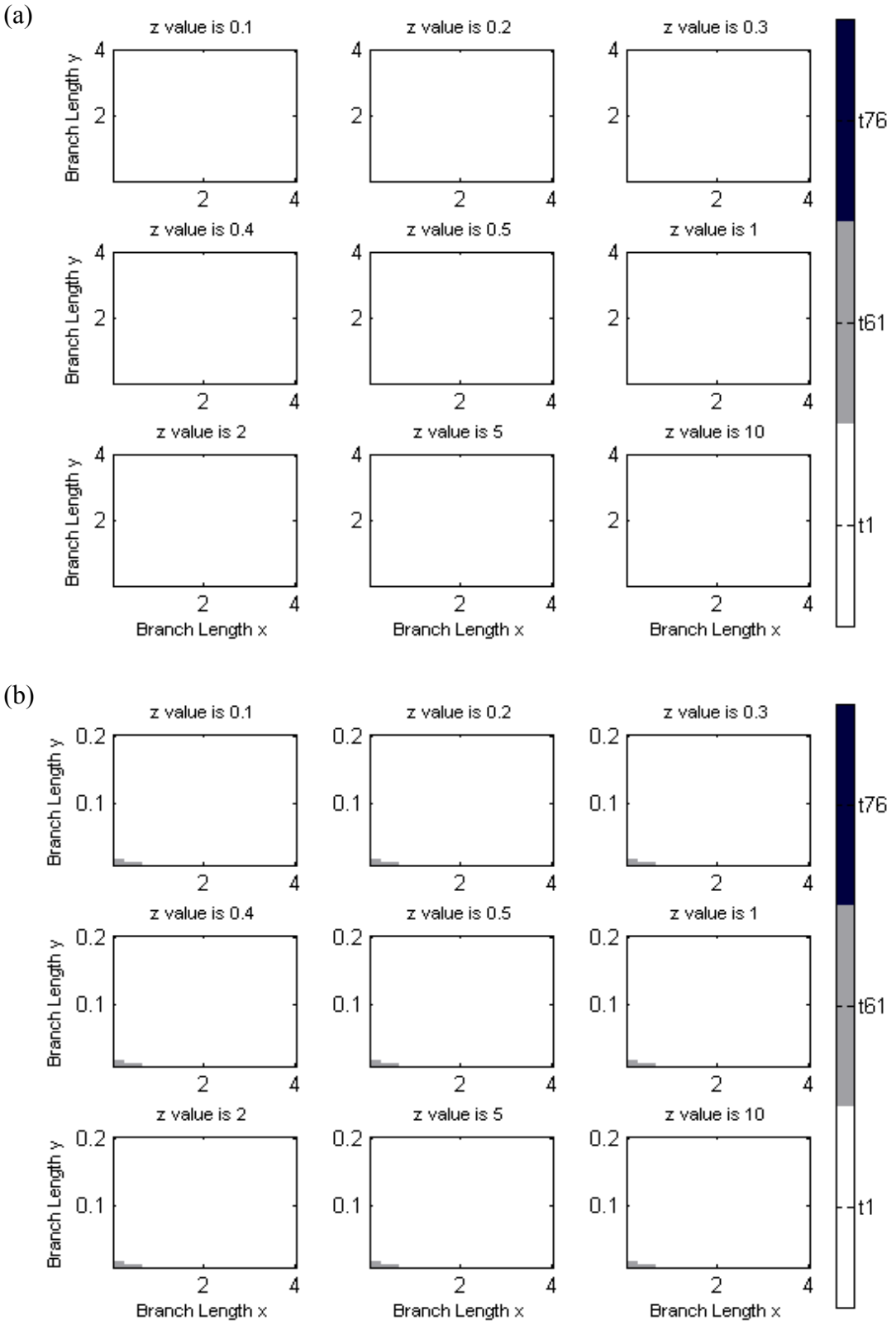


Figure 47. Pruning 1 taxa randomly
under the true species tree topology t1=((((A,B),C),D),E).
Plot (b) is a zoom in of plot (a).

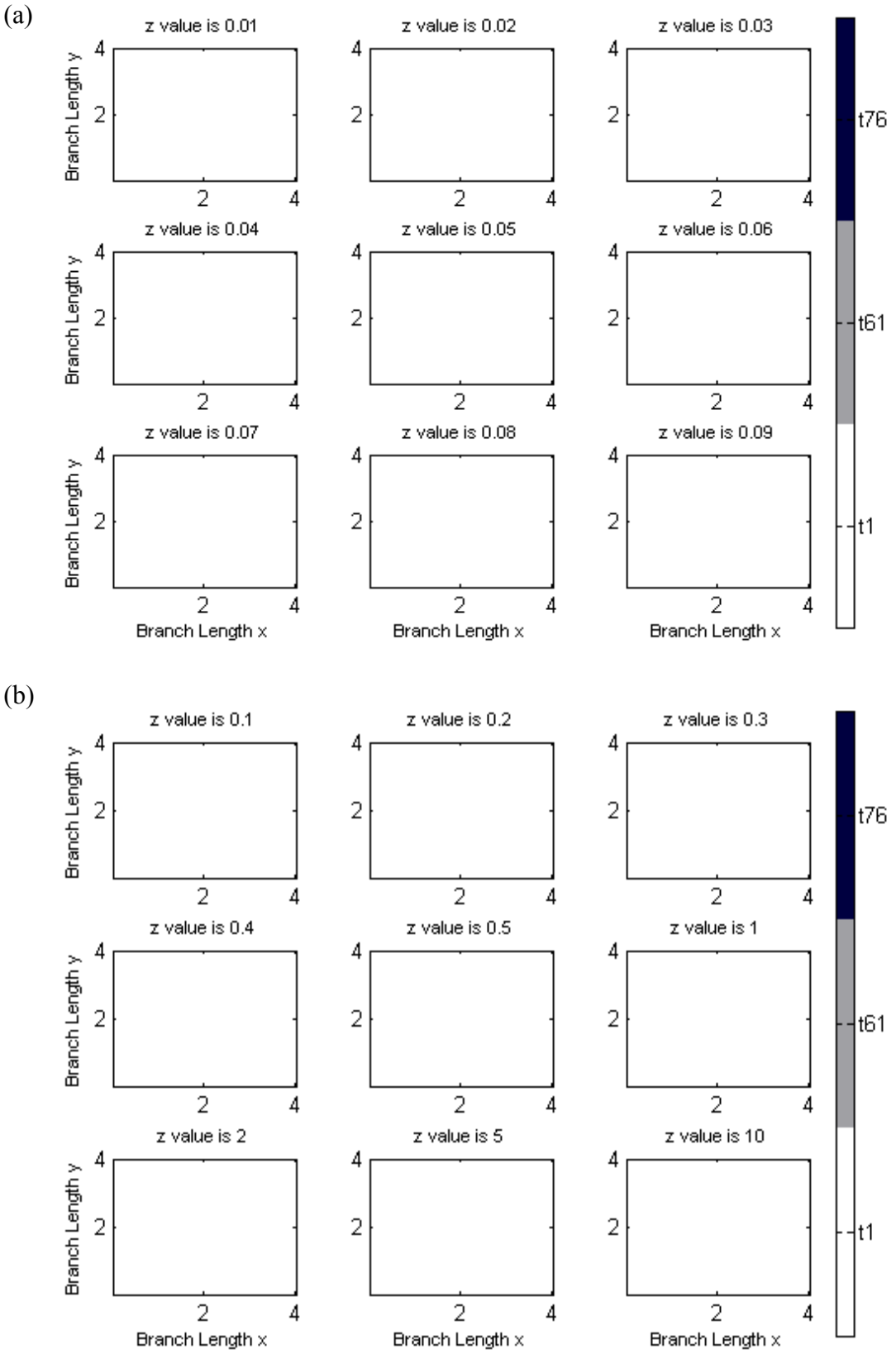


Figure 48. Pruning 2 taxa randomly
under the true species tree topology t1=((((A,B),C),D),E).

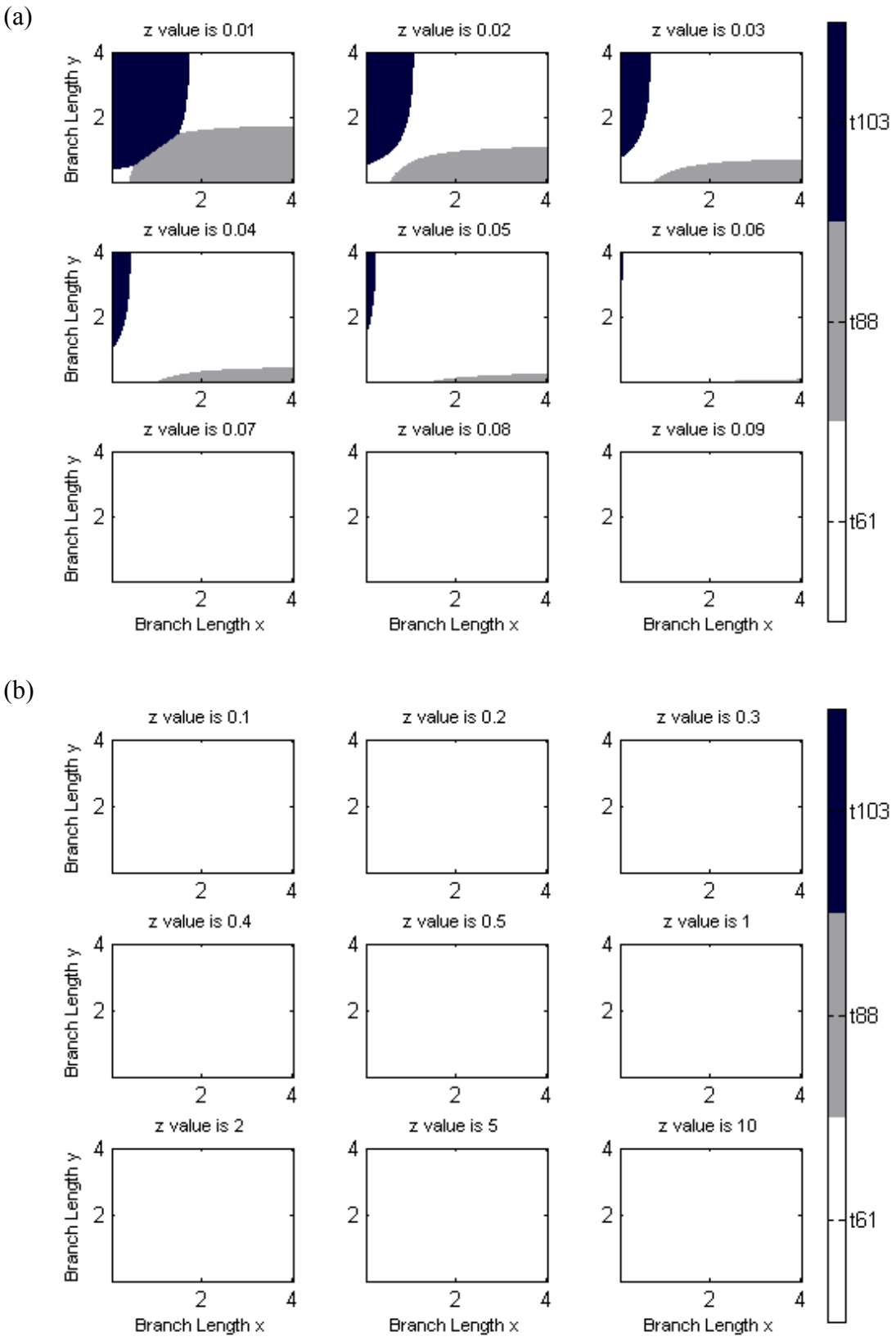


Figure 49. No pruning
under the true species tree topology t61=(((A,B),(C,D)),E).

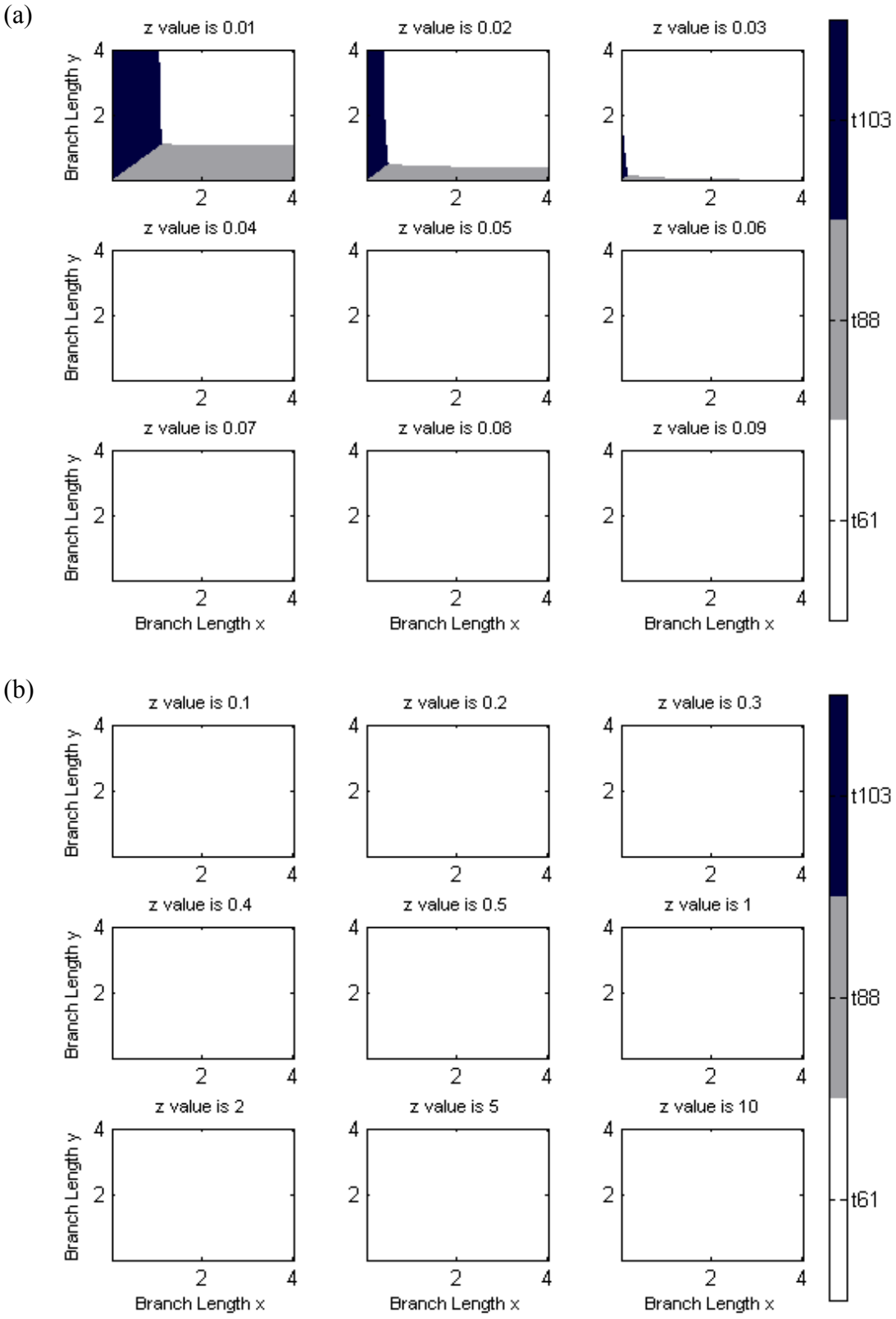


Figure 50. Pruning 1 taxon randomly
under the true species tree topology t61=(((A,B),(C,D)),E).

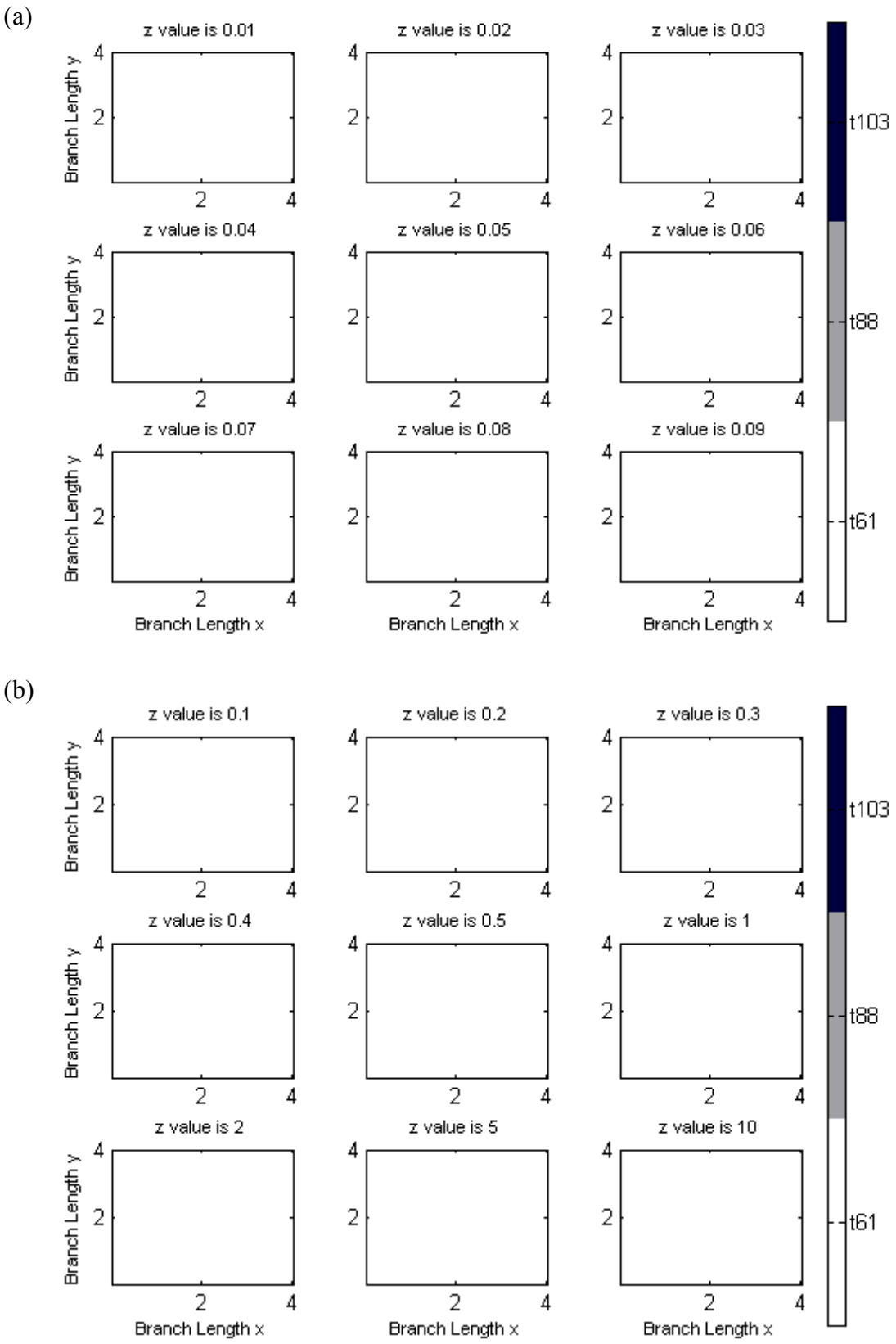


Figure 51. Pruning 2 taxa randomly
under the true species tree topology t61=(((A,B),(C,D)),E).

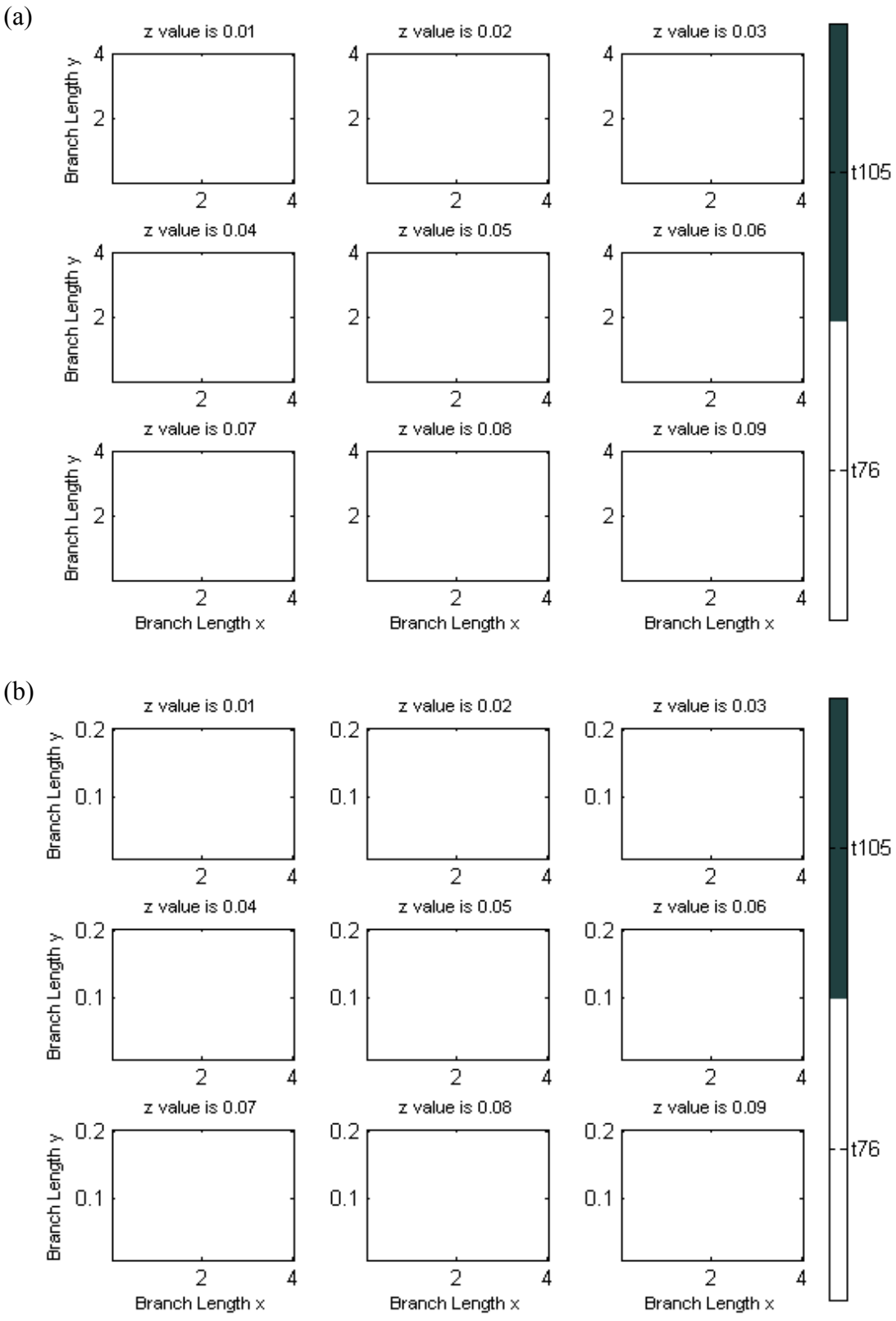


Figure 52. No pruning
under the true species tree topology t76=(((A,B),C),(D,E)).
Plot (b) is a zoom in of plot (a).

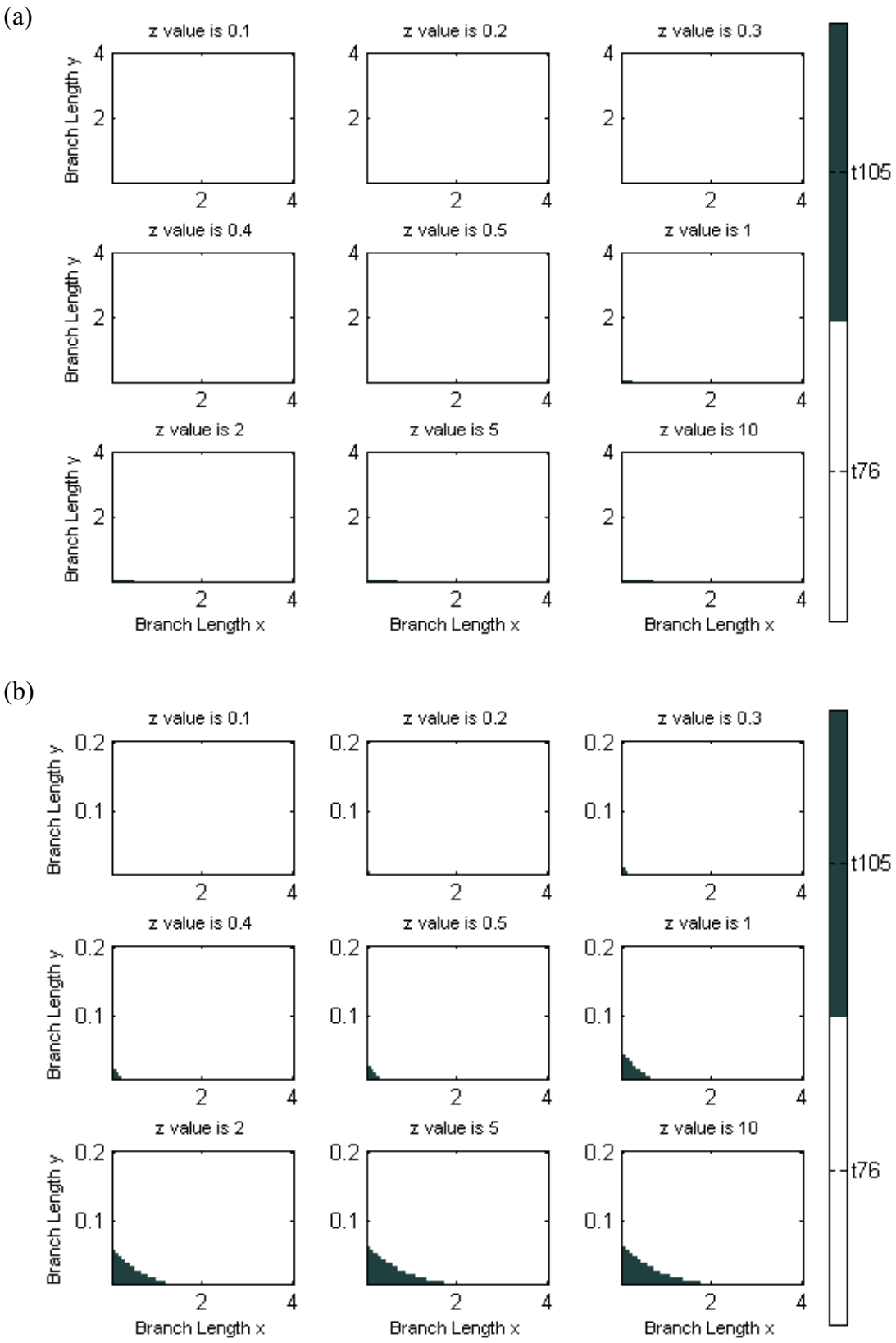
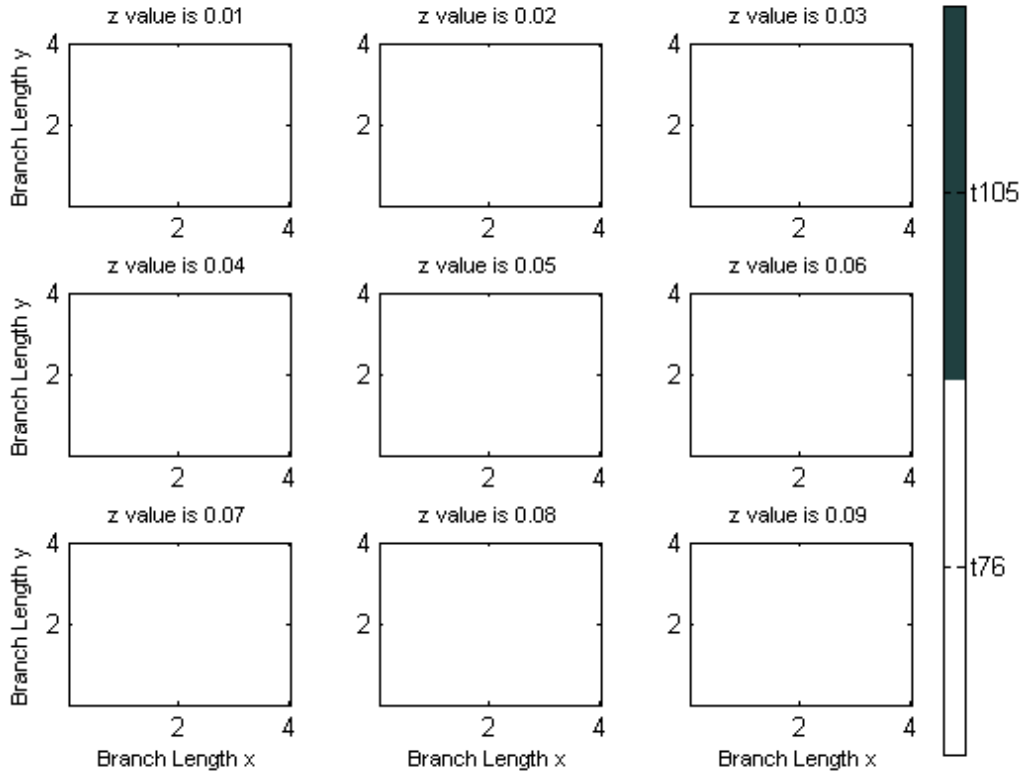


Figure 53. No pruning
under the true species tree topology t76=(((A,B),C),(D,E)).
Plot (b) is a zoom in of plot (a).

(a)



(b)

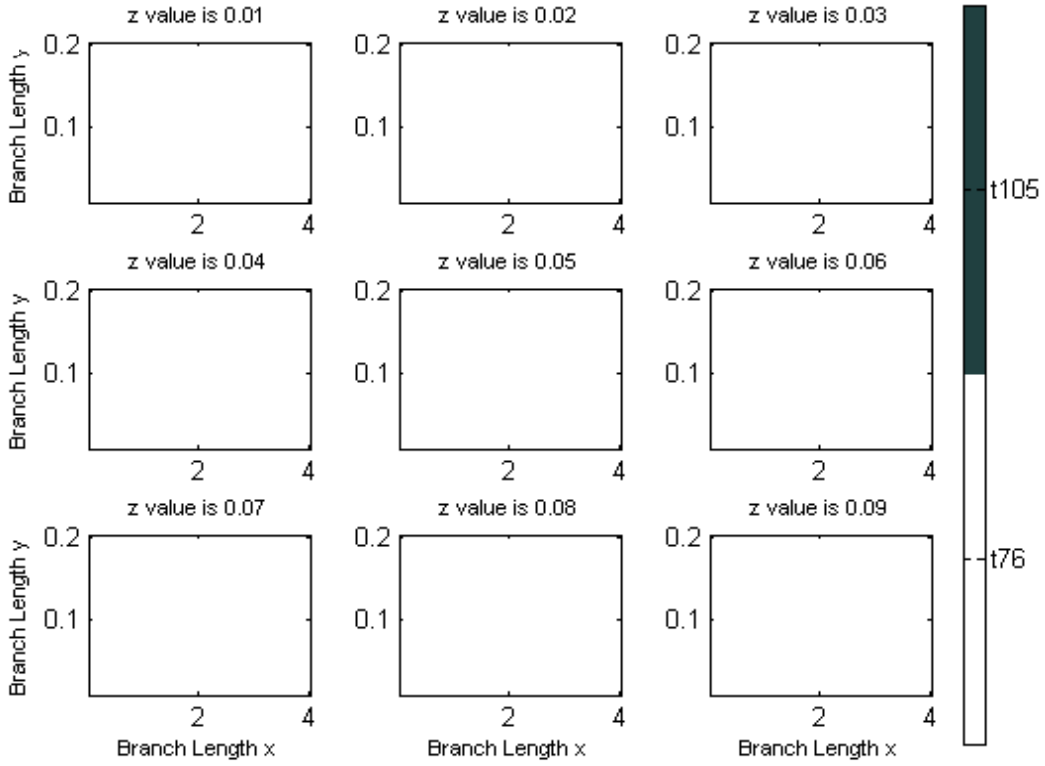
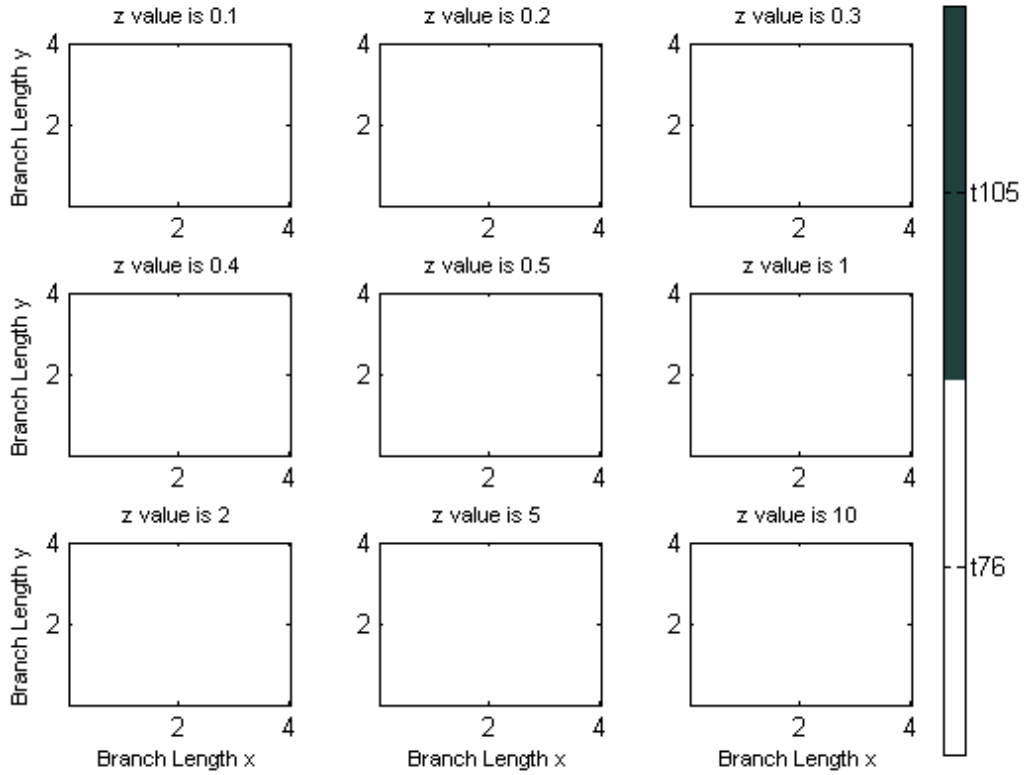


Figure 54. Pruning 1 taxon randomly under the true species tree topology t76=(((A,B),C),(D,E)). Plot (b) is a zoom in of plot (a).

(a)



(b)

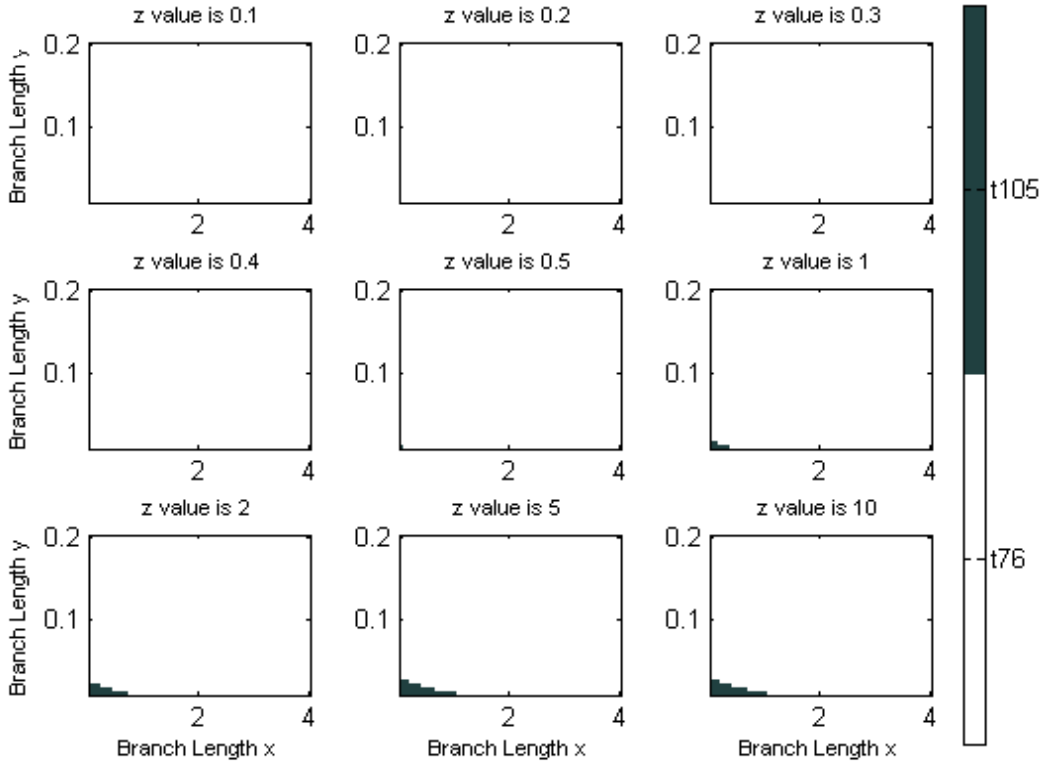
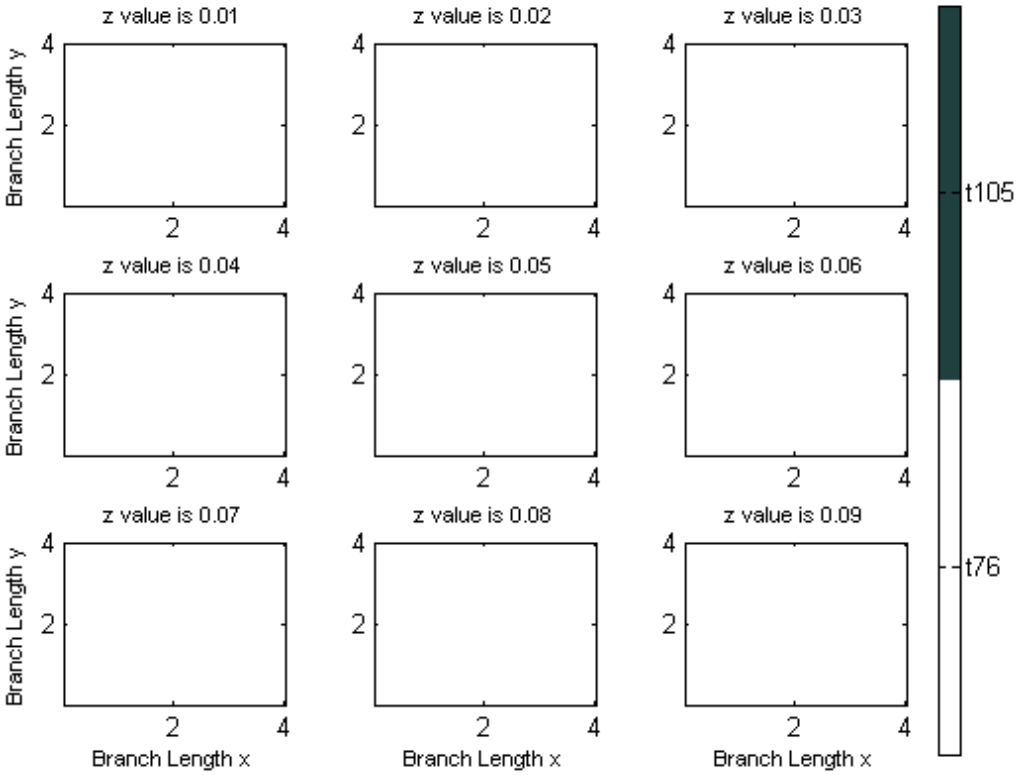


Figure 55. Pruning 1 taxon randomly under the true species tree topology t76=(((A,B),C),(D,E)). Plot (b) is a zoom in of plot (a).

(a)



(b)

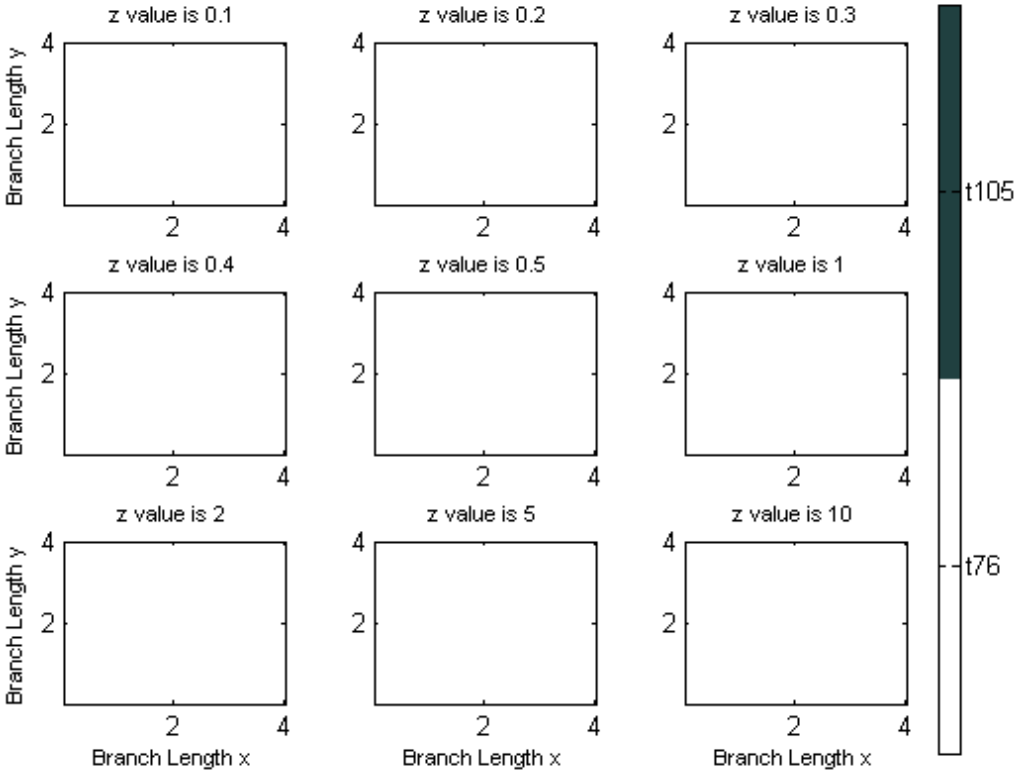


Figure 56. Pruning 2 taxa randomly
under the true species tree topology t76=(((A,B),C),(D,E)).

5 Summary

A major goal of evolutionary biology is to gain a better insight of the tree of life, the tree of ancestor-descendant relationships for all species. This mission involves constructing an estimated species tree with gene trees estimated from many loci. If there are missing taxa in the input gene trees (i.e. missing data), a supertree method like MRP can be used. In this thesis, under the multispecies coalescent model, there were 2 main tasks: to demonstrate the performance of MRP for various simulation settings and to illustrate the implementation of MRPAST.

The simulation results show that a matching estimated species tree topology is not always returned by MRP even as the sample size approaches infinity. In other words, MRP is not statistically consistent. However, the pruning schemes and mutation can improve the performance of MRP so that the matching estimated species tree topology is returned more often in some cases. In particular, the results suggest that if all the input gene trees are randomly pruned to form a rooted triple, the topology of the resulting estimated species tree always matches the true species tree topology.

We also show that with infinitely many simulated gene trees and under the consensus setting, MRP is equivalent to the greedy consensus for both 4-taxon species trees ((A,B),(C,D)) and (((A,B),C),D). That is, the same estimated species tree topology is returned by both methods for the same combination of branch lengths (x, y) in the

consensus setting.

One possible extension of this project is to study the performance of MRP for species tree with more than 20 taxa with simulation. Another possible extension would be to develop the MRPAST further in order to cover the mutation case. It is also a potential extension to compare MRP to the greedy consensus for more than 4-taxon species tree. In particular, we found an example where MRPAST and the asymptotic consensus tree are not equivalent for 5 taxa although they are the same for 4 taxa.

Appendix

20-taxon species trees with branch lengths and outgroup taxon U used in Chapter 3 simulation and supplementary results.

Tree 1 (((Q:6.396447897020334,(((N:0.47010465562691783,L:0.4701046556269
1783):3.614644959946449,(O:2.6264094037294226,((R:0.2786286106991
532,T:0.2786286106991532):1.668236194208755,P:1.946864804907908):
0.6795445988215144):1.4583402118439446):1.5934183637088024,((G:2.
9292160815797583,(K:2.8807037868579846,((B:1.7622408670415182,((
A:0.08682477043347289,C:0.08682477043347289):0.0991499774741333,
S:0.1859747479076062):1.576266119133912):0.7829627136071909,I:2.54
52035806487094):0.33550020620927573):0.04851229472177326):0.5815
058395389201,((E:1.1313077023505673,H:1.1313077023505673):0.3084
690427763763,M:1.4397767451269439):2.070945175991734):2.16744605
8163491):0.7182799177381661):3.603552102979663,(D:4.686678839603
5615,(F:2.049862635208939,J:2.049862635208939):2.6368162043946217
):5.313321160396438):50,U:60);

Tree 2 (((T:4.802851565039117,(H:1.9285172773150248,(P:1.922696394110775
5,R:1.9226963941107755):0.005820883204249125):2.8743342877240927
):2.4640164442179127,((C:3.493616785263557,B:3.493616785263557):3.
469745618478964,(((D:4.230452214490377,(O:2.432131075466071,K:2.4
32131075466071):1.798321139024306):2.3954796890320336,(((S:2.1365
267794620495,J:2.1365267794620495):0.6567784205535473,F:2.7933052
000155967):3.826830542643562,(Q:2.570128952536905,(I:1.8273704475
813914,N:1.8273704475813914):0.742758504955133):4.0500067901222
54):0.005796160863251664):0.21673499338613383,((L:0.727015677465
0244,E:0.7270156774650244):2.405118353170957,A:3.132134030635981
):3.710532866272563):0.12069550683397684):0.3035056055145099):2.7
331319907429688,(M:1.090383352276004,G:1.090383352276004):8.9096
16647723993):50,U:60);

Tree 3 (((M:1.1174805336553957,(S:1.1012512470208167,H:1.10125124702081
67):0.01622928663457929):6.62928691862423,(((F:3.3231665966244184,
(P:0.163048183215133,R:0.163048183215133):3.1601184134092852):3.5
971442581777398,((I:0.3513221208666479,A:0.3513221208666479):0.35

945833789204756,J:0.7107804587586953):6.209530396043463):0.594387
2404881059,(B:2.61362603709842,O:2.61362603709842):4.90107205819
1843):0.2320693569893616 5):2.253232547720373,(((D:3.9651289542947
71,(((G:1.9847460915430288,Q:1.9847460915430288):1.11521716079843
2,N:3.0999632523414604):0.1251661099552976,(C:2.206782282055621,
L:2.206782282055621):1.0183470802411367):0.7399995919980134):0.87
58812367696035,T:4.841010191064375):0.923813499554089,(K:2.30330
3190117338,E:2.303303190117338):3.461520500501125):4.23517630938
15356):50,U:60);

Tree 4 ((A:9.999999999999998,((G:6.586567444586015,T:6.586567444586015):
2.657134720448915,(((P:0.0016488808254750702,H:0.001648880825475
0702):2.6881482731500554,((R:1.619804800093447,(I:1.60011461879008
37,O:1.600114618790083 7):0.01969018130336369 6):0.666284065140002
,(S:0.7897903562631157,D:0.7897903562631157):1.4962985089703338):
0.4037082887420808):2.812446570699692 4,(((J:0.9491526197050424,N:
0.9491526197050424):3.6944829985237333,(Q:2.3965123998650553,B:2.
3965123998650553):2.24712321836372):0.4695549215371633,E:5.11319
0539765938):0.289212963687295,((F:3.797203951110874,(K:1.81811154
3149181,C:1.818111543149181):1.9790924079616938):0.4890146660037
834,(L:2.021966066035451,M:2.021966066035451):2.264252551079207):
1.116184886338576):0.09984022122198835):3.7414584403597084):0.756
2978349650682):50,U:60);

Tree 5 ((((O:3.6517682163652463,R:3.6517682163652463):1.4839381794901547
,(F:1.1702454826575965,((C:0.04813045465328715,L:0.04813045465328
715):1.1184332931063976,H:1.166563747759684 9):0.0036817348979118
207):3.9654609131978042):2.247612287676637,((S:2.852283761427436,(
(M:0.7949597030935436,Q:0.7949597030935436):1.430375766425931,(J:
1.5046456473497751,I:1.5046456473497751):0.7206898221696996):0.62
69482891907962 1):4.298438439283786,(D:3.3081370020357923,((B:0.957
3517330582655,E:0.9573517330582655):2.3384515600942493,N:3.29580
32931525145):0.012333708883277577):3.8425851986754305):0.2325964
8282081458):2.6166813164679623,((G:0.1161797579686813 5,A:0.11617
975796868135):9.106939051262831,((T:1.5440277613535596,K:1.544027
7613535596):0.7453893231660852,P:2.289417084519645):6.9337017247
118675):0.776881190768488):50,U:60);

Tree 6 (((((L:0.012827022148860532,Q:0.012827022148860532):3.31283631153
5691,((F:2.8004110959389448,H:2.8004110959389448):0.5197494320177
148,((E:0.7865864384140449,K:0.7865864384140449):2.5230713934205
88,((A:0.9532290816487597,G:0.9532290816487597):1.29889534880635

42,R:2.252124430455114):1.057533401379519):0.010502696122026553):0.005502805727892291):1.4419070394776228,(I:2.0099028207997875,B:2.0099028207997875):2.757667552362387):2.458183371935972,((P:2.375977097170804,(O:1.0929172670073148,(M:0.7613839331543607,J:0.7613839331543607):0.3315333338529543):1.283059830163489):4.594158532314615,((T:1.1603901121508065,N:1.1603901121508065):1.827177450568165,C:2.9875675627189713):3.982568066766447):0.25561811561272707):2.774246254901852,(D:1.4951325867946617,S:1.4951325867946617):8.504867413205337):50,U:60);

Tree 7 (((K:3.3824065355200554,O:3.3824065355200554):4.444634724311502,(((D:0.37702881304849334,F:0.37702881304849334):0.9300502159540328,N:1.307079029002526):4.937796161489638,R:6.244875190492165):1.582166069339394):2.1729587401684407,((S:5.166721725682582,((E:0.9781833920258888,Q:0.9781833920258888):1.7065279002150242,(G:2.659564290306099,(J:2.6586299372714897,(P:0.24225855747004169,I:0.24225855747004169):2.416371379801448):9.343530346092986E-4):0.025147001934814037):2.48201043344167):1.714808521702752,((C:2.145437485983847,(H:2.126987856736408,A:2.126987856736408):0.018449629247439345):0.8054422568168799,(M:1.9214405249021012,((L:0.16157175347011693,T:0.16157175347011693):1.1179292518175332,B:1.279501005288765):0.6419395196144508):1.0294392178986258):3.930650504584608):3.1184697526146645):50,U:60);

Tree 8 (((P:5.28845289818418,((S:3.0887490032248053,(I:3.065457751147678,J:3.065457751147678):0.02329125207712752):0.3877263807135393,(H:1.3349665157038613,F:1.3349665157038613):2.141508868234484):1.8119775142458343):4.71154710181582,(((B:0.3958363726009147,K:0.3958363726009147):1.0560256396006713,N:1.4518620122015862):0.7295453511313612,L:2.181407363332947):6.9026770326714,(((((T:3.6404384174852513,O:3.6404384174852513):0.772029320773687,G:4.412467738258939):2.5786832630542667,(C:5.153380277922956,((R:1.1052069704669119,Q:1.1052069704669119):2.193711848232931,D:3.2989188186998435):1.8544614592231123):1.8377707233902496):0.7422767504536996,A:7.733427751766904):0.5409744523653208,E:8.274402204132224):0.6683638394330498,M:8.942766043565276):0.14131835243907231):0.9159156039956521):50,U:60);

Tree 9 (((H:3.5465391661611734,(((R:0.2243883634649781,B:0.2243883634649781):0.8037659361307938,D:1.028154299595772):1.0127912480942873,M:2.0409455476900593):1.0553108903022097,J:3.0962564379922686):0.45028272816890497):6.4534608338388235,(((N:1.2136287569614905,G:

1.2136287569614905):4.298917698882908,((K:2.8506563200726163,L:2.8506563200726163):2.5898277231198685,C:5.440484043192487):0.07206241265191403):2.262712926610753,P:7.775259382455152):1.881849704903281,(((T:0.441409045685205,F:0.441409045685205):1.7791577657702877,Q:2.220566811455493):4.780736434171814,(A:3.2417753974223373,((I:0.039134605155767596,S:0.039134605155767596):3.171093482085578,E:3.2102280872413456):0.03154731018099152):3.7595278482049697):2.4851753425189385,O:9.486478588146246):0.17063049921218798):0.34289091264156446):50,U:60);

Tree 10 ((((D:1.6163170972997407,(((F:0.019620849389433002,K:0.019620849389433002):0.13898991270413016,S:0.15861076209356317):0.11398544375011697,P:0.27259620584368016):0.12169723936239996,R:0.3942934452060802):1.222023652093661):3.292678617934042,(((O:2.842324414716848,((J:1.5065719321593305,(M:0.4996525247162146,T:0.4996525247162146):1.0069194074431165):1.2860309154285428,G:2.7926028475878732):0.04972156712897484):0.5070777420883522,E:3.3494021568052004):0.8791164029903944,I:4.228518559795595):0.5064694493730965,((B:3.9103567522822957,(C:0.5463472298253979,A:0.5463472298253979):3.364009522456898):0.7243869806775107,Q:4.634743732959807):0.10024427620888467):0.17400770606509192):3.28375528284326,N:8.192750998077043):1.8072490019229563,(L:0.749499867505291,H:0.749499867505291):9.25050013249471):50,U:60);

Tree 11 (((((N:0.8687758944484617,(A:0.040265376280712396,F:0.040265376280712396):0.8285105181677493):1.278161061565935,(I:2.125027363489056,(E:0.3456754087465365,D:0.3456754087465365):1.7793519547425194):0.0219095925253408):4.0083956761493456,(O:1.635082054953186,P:1.635082054953186):4.520250577210556):1.8511062064417896,(S:5.013907460049822,(C:0.10864613099676307,(J:0.10568871287285828,M:0.10568871287285828):0.0029574181239047878):4.905261329053058):2.9925313785557113):1.993561161394467,((B:4.343506602952826,Q:4.343506602952826):1.1651086726862345,((G:0.6075179531683955,T:0.6075179531683955):3.4992740810041534,(((H:0.43232636450449147,K:0.43232636450449147):1.5518112515091413,L:1.9841376160136328):2.0900864705825075,R:4.07422408659614):0.032567947576408):1.4018232414665115):4.491384724360939):50,U:60);

Tree 12 ((((R:2.049476917223095,(K:1.5236168620163253,C:1.5236168620163253):0.5258600552067705):2.5407981037139065,F:4.5902750209370025):2.5142587441626385,(((H:1.2003661212798005,E:1.2003661212798005):0.38587816811582254,M:1.586244289395623):2.564686090178316,(S:0.0

20869799310792424,O:0.020869799310792424):4.130060580263146):2.6691354321230922,(((D:0.9447220808187581,(Q:0.5340460717472297,L:0.5340460717472297):0.4106760090715284):2.9162918530003568,(B:1.9585988274763177,N:1.9585988274763177):1.902415106342797):2.9545665996742208,I:6.815580533493338):0.004485278203695116):0.2844679534026099):2.895466234900356,(((T:0.1752360353889561 9,(G:0.1686285960377976,J:0.1686285960377976):0.006607439351158572):0.3741523910807516,A:0.5493884264697078):2.6126335192120136,P:3.162021945681721):6.837978054318275):50,U:60);

Tree 13 ((((F:2.1502539033237027,(A:0.8966775135025297,C:0.8966775135025297):1.2535763898211723):2.623374895000622,K:4.773628798324325):1.717743265129889,(H:1.5590676880487537,((T:0.0011090275462978625,O:0.0011090275462978625):1.0807857628730764,(S:0.4045859771152553,E:0.4045859771152553):0.677308813304119):0.47717289762937964):4.93230437540546):3.508627936545785,((M:3.1653593189514804,(((((J:0.1853580404095614,I:0.1853580404095614):0.29288450956179496,Q:0.4782425499713564):0.21577699580526083,((P:0.13314857447484954,D:0.13314857447484954):0.5484214224661588,B:0.6815699969410083):0.01244954883560891):0.46297753051990526,L:1.1569970762965223):0.19999620675423363,N:1.356993283050756):0.7062361299384579,G:2.0632294129892146):1.1021299059622662):0.7557102563372711,R:3.921069575288752):6.078930424711247):50,U:60);

Tree 14 ((((J:2.1282591958948016,(O:2.1070859386887353,(P:0.5788744760218603,R:0.5788744760218603):1.528211462666875):0.02117325720606596):1.4223146502550528,(F:1.7206520397498086,E:1.7206520397498086):1.829921806400045):1.3672647224498018,(((C:1.8633300091723355,(K:1.6930590859650578,H:1.6930590859650578):0.17027092320727796):2.802734850662005,Q:4.666064859834341):0.16630369304696715,S:4.832368552881309):0.08547001571834781):5.082161431400342,(((L:0.6733030552249427,I:0.6733030552249427):7.462351464983974,(A:0.2872609122433474,D:0.2872609122433474):7.848393607965569):1.0510411204988117,((M:1.323802609398045,T:1.323802609398045):1.796395062942878,(B:0.14539902503169252,G:0.14539902503169252):2.9747986473092305):6.0664979683668045):0.7289402866333777,N:9.915635927341107):0.08436407265889187):50,U:60);

Tree 15 (((H:6.4938105860469335,((N:5.679876958022631,(S:4.80927923203247,(P:4.753826247537515,D:4.753826247537515):0.05545298449495615):0.8705977259901607):0.6744971880011386,(((R:0.09994579194525016,G:0.09994579194525016):3.925821199444271,L:4.025766913895215):1.31

11434178963537,((I:2.2984350118299894,O:2.2984350118299894):3.015
9973074960598,J:5.314432319326048):0.022478089959825684):1.017463
7367378958):0.13943644002316447):3.5061894139530643,((K:1.1537517
223861484,((Q:0.8814212117977878,C:0.8814212117977878):0.26153235
24197475,B:1.142953564217535):0.010798158168613528):1.6191469551
774564,(A:0.9906495330362384,(((M:0.7294549895514751,F:0.72945498
95514751):0.07998596374169213,E:0.8094409532931672):0.1652484280
460107,T:0.9746893813391778):0.0159601516970605 4):1.782249144527
3667):7.227101322436395):50,U:60);

Tree 16 ((((H:1.8121712070212843,B:1.8121712070212843):0.9722915842044316
,(((J:0.1348896486927485,G:0.1348896486927485):1.2127965202397488,
T:1.3476861689324973):0.3546190404048792,K:1.7023052093373765):1.
0241138763390119,((R:0.5641748149784542,((D:0.28418585619081443,
C:0.28418585619081443):0.18995586404434897,E:0.4741417202351634)
:0.0900330947432909):0.6763646021902031,(M:1.225899228711816,P:1.
225899228711816):0.01464018845684149):1.4858796685077307):0.0580
43705549327904):2.0818386245813687,(O:0.6991925004053641,N:0.699
1925004053641):4.167108915401721):5.133698584192914,((A:0.9509244
936121762,F:0.9509244936121762):3.4273150237914374,((L:0.30958090
87794897,Q:0.3095809087794897):1.1939765140762961,(I:0.4878434163
0104205,S:0.48784341630104205):1.0157140065547439):2.87468209454
7828):5.621760482596385):50,U:60);

Tree 17 (((J:3.433574823567465,((Q:1.0955256277315508,N:1.095525627731550
8):2.009021544123965,(D:0.9602739261935831,G:0.9602739261935831):
2.144273245661933):0.32902765171194914):3.508413080908794,((A:1.9
270343454192878,((T:0.08067957933793155,S:0.08067957933793155):0.
11058145164948639,F:0.19126103098741795):1.7357733144318699):0.7
207677189528493,K:2.6478020643721374):4.294185840104122):3.05801
20955237406,(((E:1.2898824078642623,R:1.2898824078642623):3.30851
80517616424,(C:0.7092959976984125,H:0.7092959976984125):3.889104
4619274924):1.2112156911334802,((L:2.306412668691293,M:2.3064126
68691293):0.9211175910388625,P:3.2275302597301554):2.58208589102
923):3.059451998534483,(B:0.28345460491462354,(I:0.28223807303251
724,O:0.28223807303251724):0.0012165318821063115):8.585613544379
243):1.1309318507061312):50,U:60);

Tree 18 ((((C:1.5830143027526167,((T:0.4828790014800336,K:0.4828790014800
336):0.06992041503429877,O:0.5527994165143324):1.030214886238284
3):2.919700198782594,((D:1.2476209263164137,E:1.2476209263164137)
:2.399384710259879,(I:2.442691356674705,L:2.442691356674705):1.204

314279901587):0.8557088649589182):2.547432225216877,((S:0.6236309928369856,G:0.6236309928369856):2.01406594636513,(B:1.151716952606971 5,H:1.1517169526069715):1.485979986595144):4.41244978754997 15):2.949853273247913,((Q:3.7695726491131776,R:3.7695726491131776):2.6753198443843775,(N:3.2340188415991893,((A:0.1510063745181427,P:0.1510063745181427):3.0240786564411186,(M:0.3360897723876945,(J:0.3275530507540198,F:0.3275530507540198):0.0085367216336747):2.8389952585715665):0.0589338106399282 2):3.210873651898366):3.5551075065024444):50,U:60);

Tree 19 ((((L:0.09209086933008942,H:0.09209086933008942):3.7408337826910447,((J:0.6897801417298858,(M:0.6067005782263125,C:0.6067005782263125):0.08307956350357318):1.381201629504157,E:2.0709817712340426):1.7619428807870914):2.5114482616551705,T:6.344372913676307):3.6556270863236935,(((R:1.5014815856287964,(A:0.6192796031501525,P:0.6192796031501525):0.882201982478644):1.6635169033148134,((Q:0.9159648302061507,(S:0.9086715695021907,N:0.9086715695021907):0.007293260703959998):2.200752621500643,D:3.116717451706794):0.04828103723681568):2.15500493678201,G:5.320003425725622):3.626176384609279,(((O:1.0917989600385605,(I:1.0761479632266886,B:1.0761479632266886):0.015650996811871654):1.1793058142808588,K:2.2711047743194195):0.44209232718295094,F:2.7131971015023697):6.232982708832529):1.0538201896650994):50,U:60);

Tree 20 (((((((R:0.4461994946608132,O:0.4461994946608132):0.8371192997215148,C:1.283318794382328):0.9435415231456397,T:2.2268603175279673):1.5273006229725499,J:3.754160940500517):1.5355068793291464,((L:0.2925402448104481,A:0.2925402448104481):3.357045844288835,(I:3.631015979094741,S:3.631015979094741):0.01857011000454248):1.6400817307303803):0.6290057060803992,D:5.918673525910063):1.2807599669192689,(((M:0.021343424281856684,F:0.021343424281856684):1.0226228434660802,N:1.0439662677479369):0.9202049200542237,(((P:0.23168786795325108,(G:0.012071601904334279,E:0.012071601904334279):0.21961626604891682):0.7476930489496685,H:0.9793809169029197):0.28398826211452943,(B:1.0254566153753495,K:1.0254566153753495):0.23791256364209956):0.7008020087847114):5.235262305027171):2.800566507170669,Q:9.999999999999996):50,U:60);

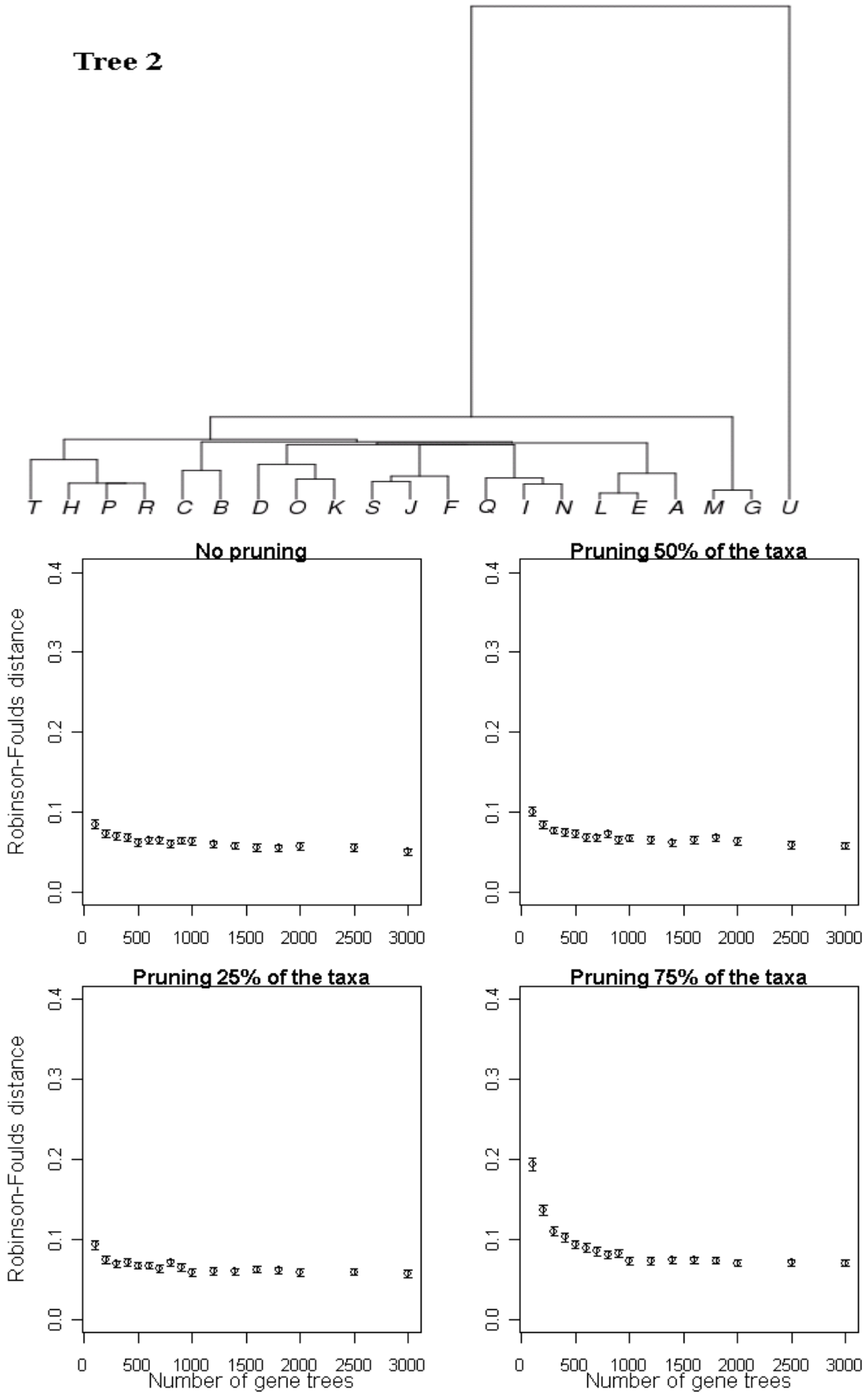


Figure 57. Tree 2 with simulated gene trees.

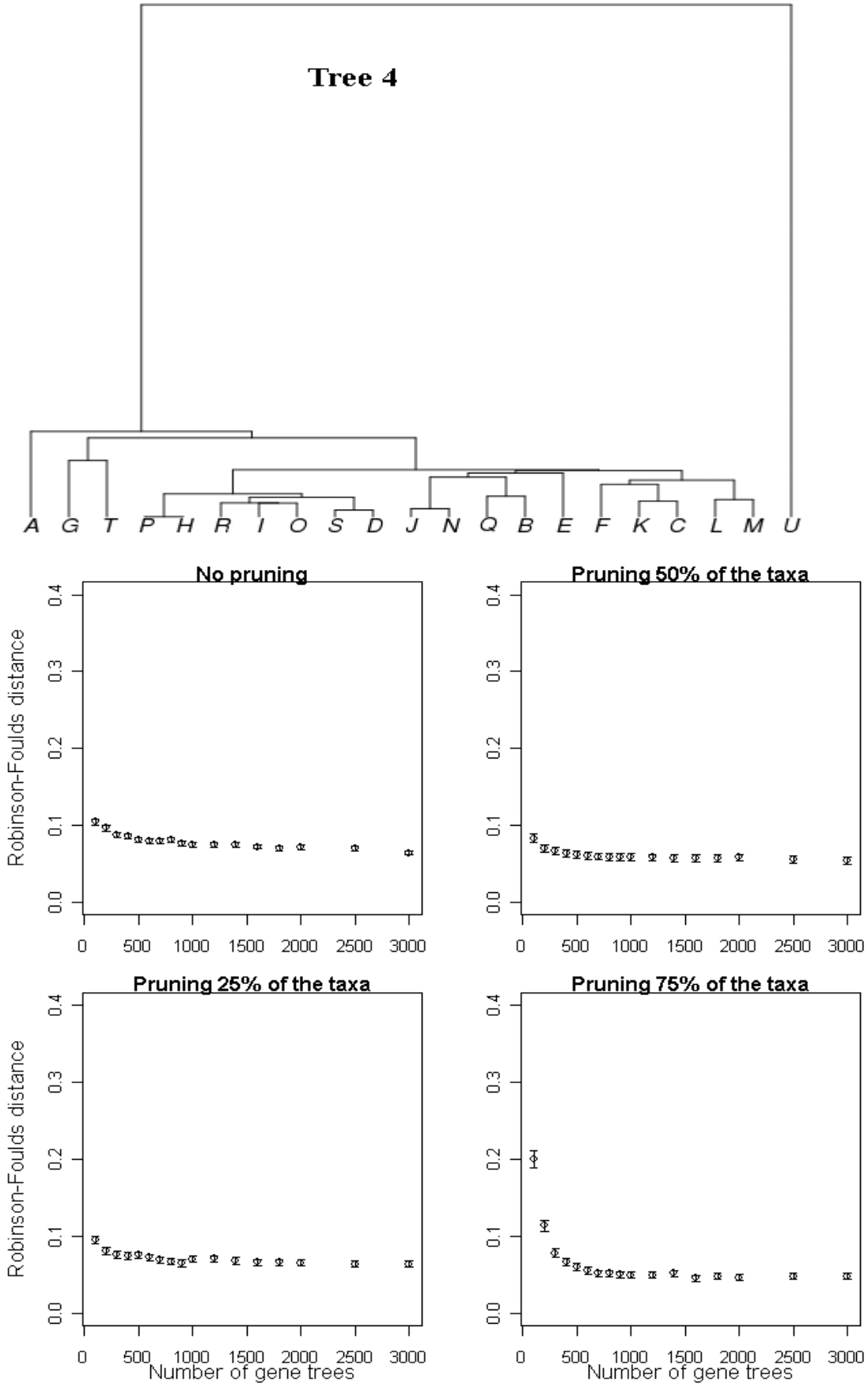


Figure 58. Tree 4 with simulated gene trees.

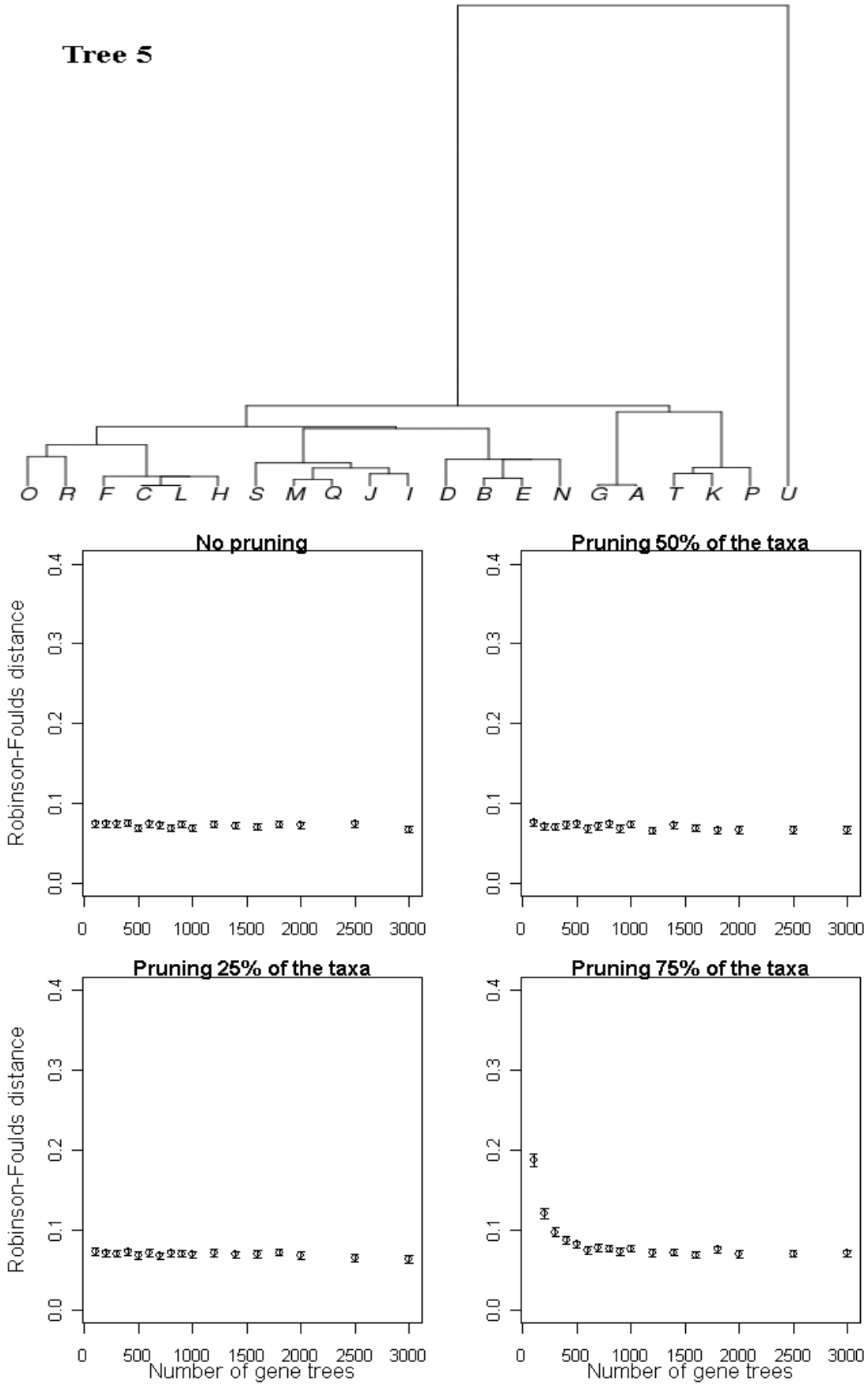


Figure 59. Tree 5 with simulated gene trees.

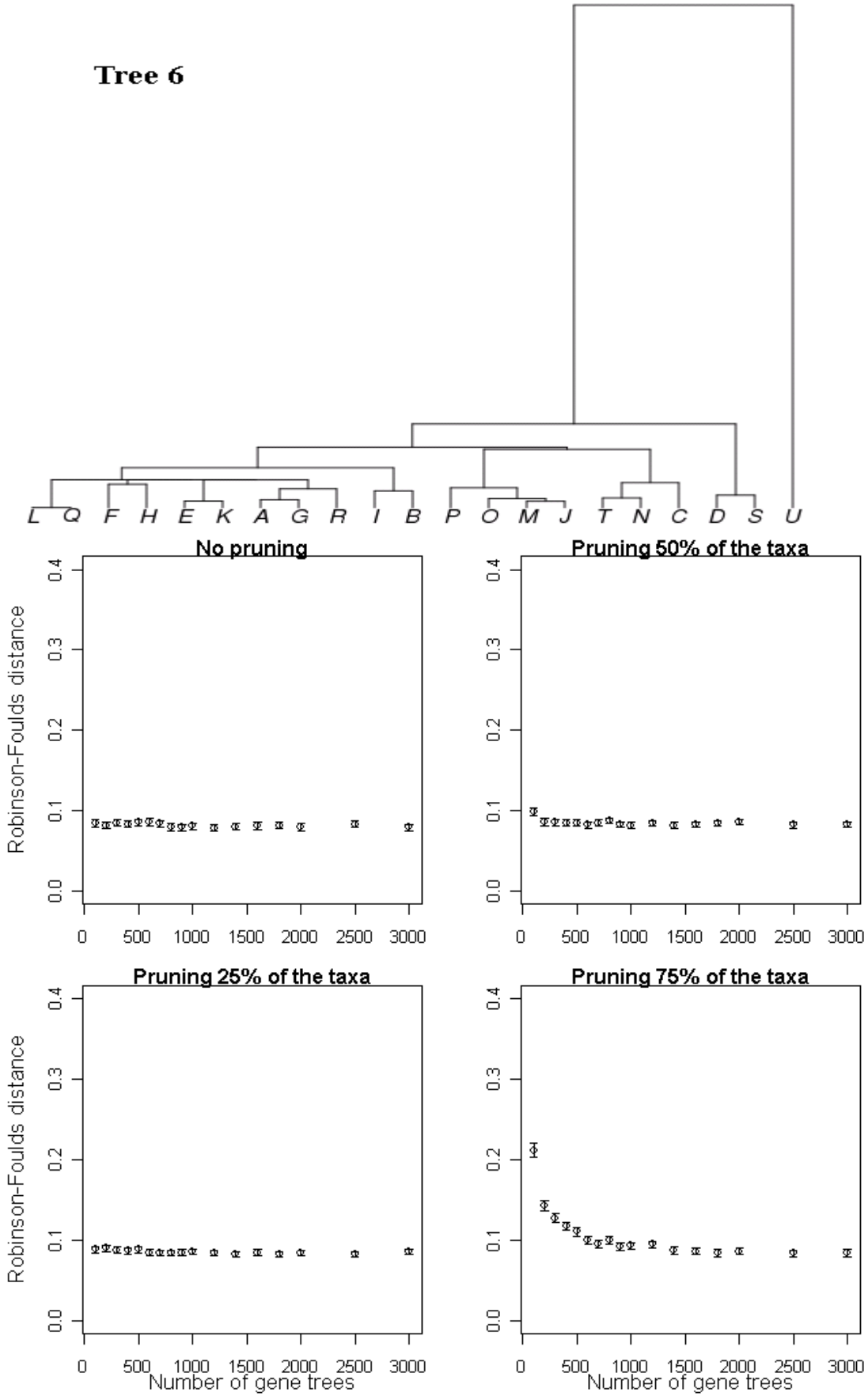


Figure 60. Tree 6 with simulated gene trees.

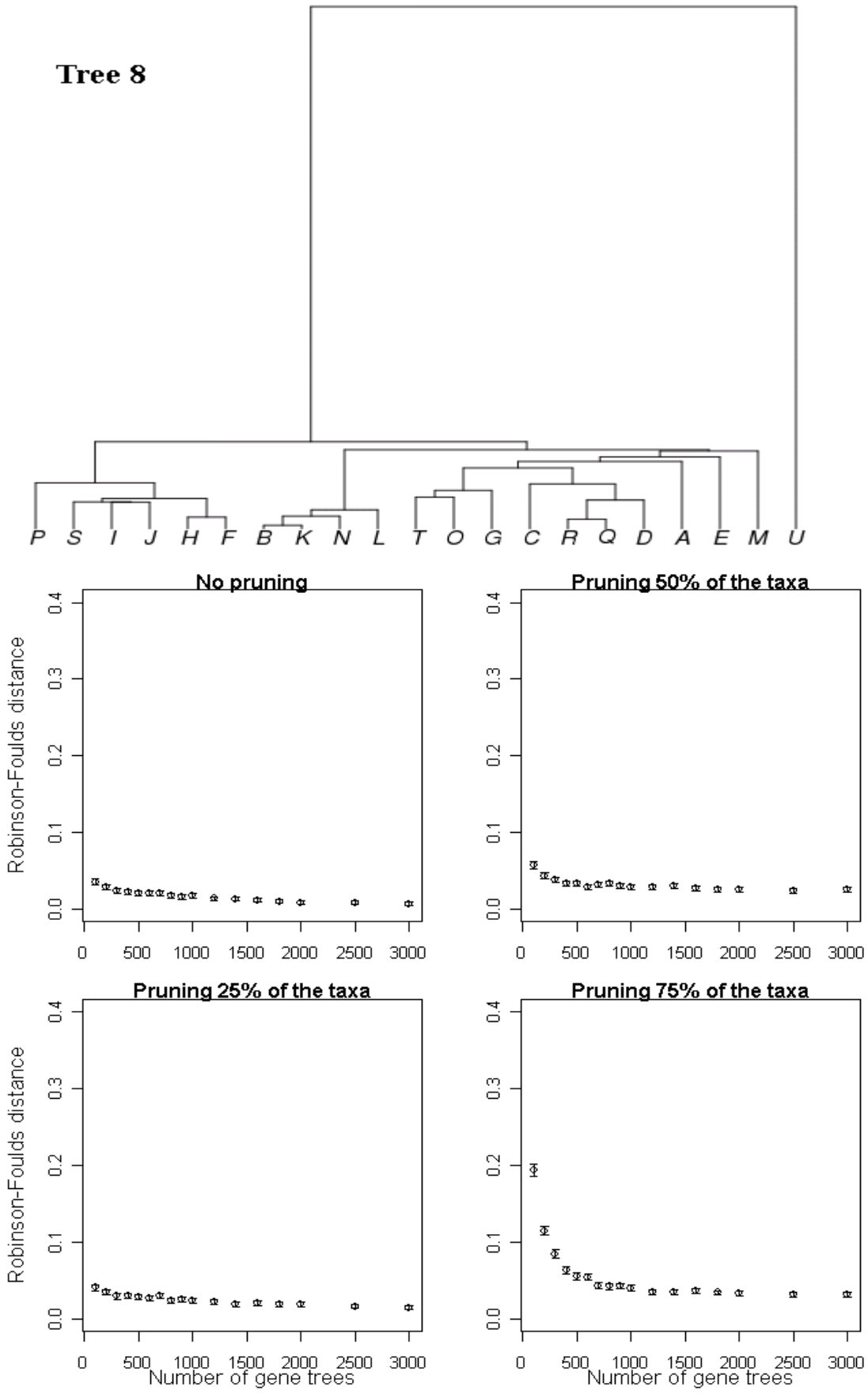


Figure 61. Tree 8 with simulated gene trees.

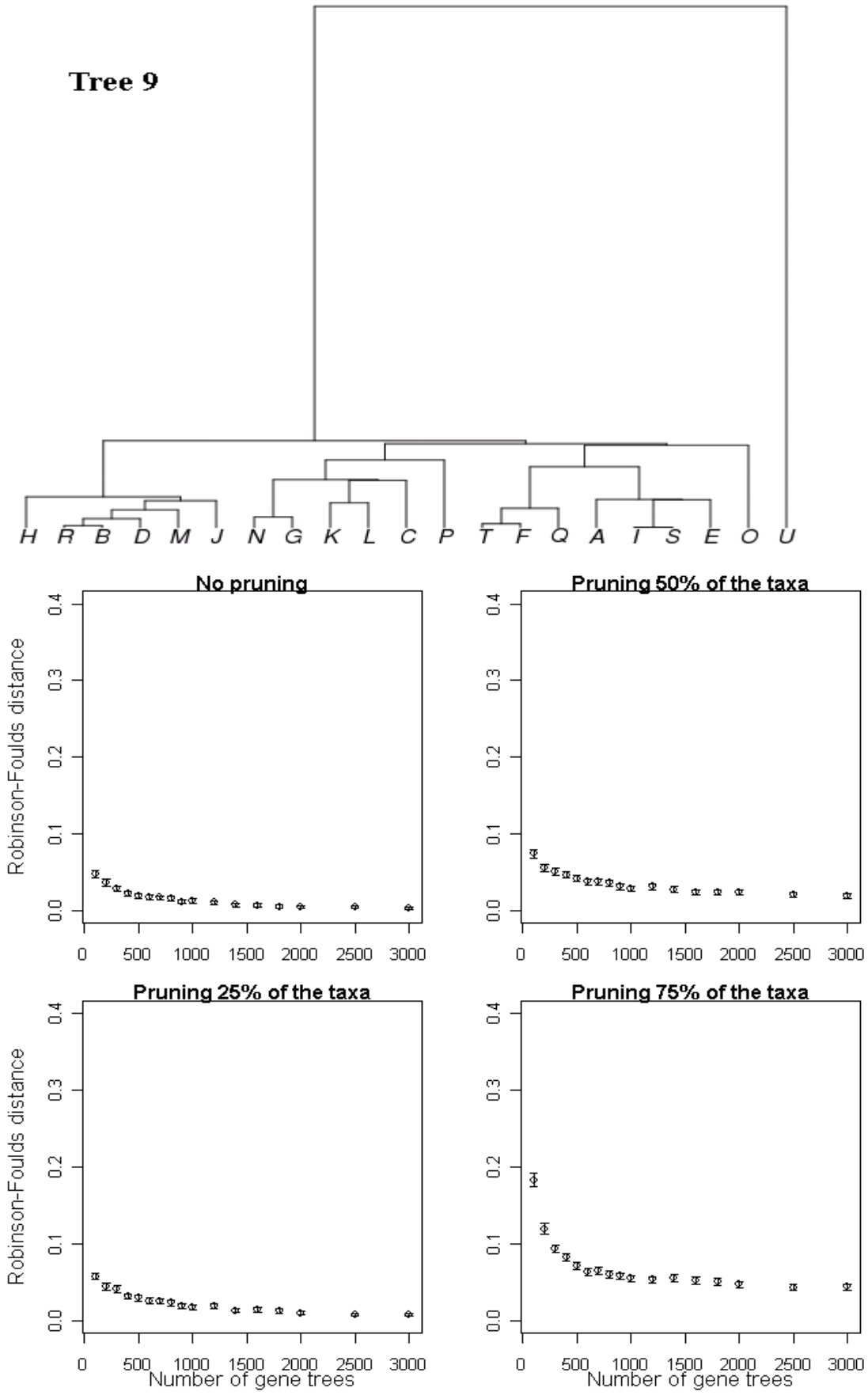


Figure 62. Tree 9 with simulated gene trees.

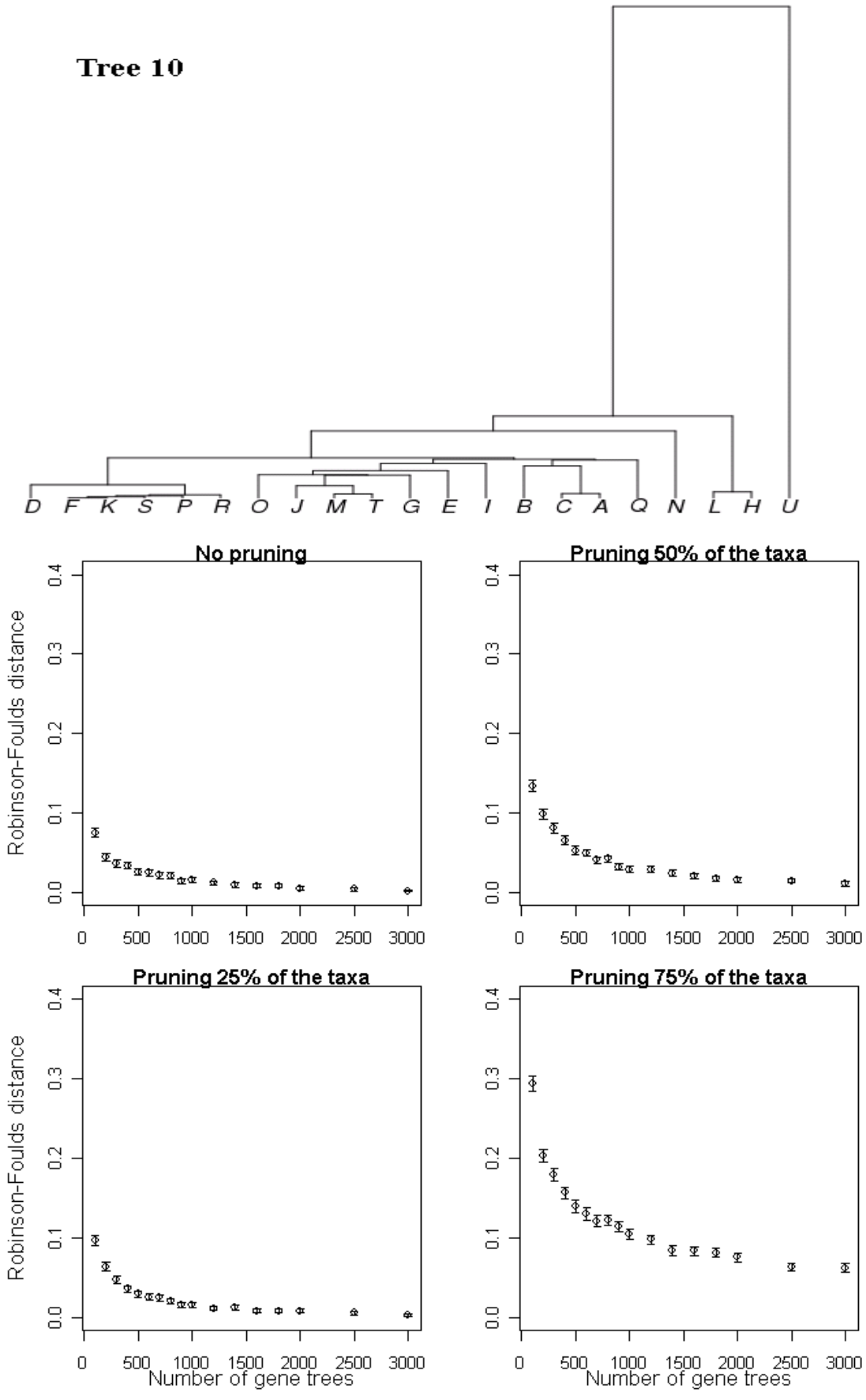


Figure 63. Tree 10 with simulated gene trees.

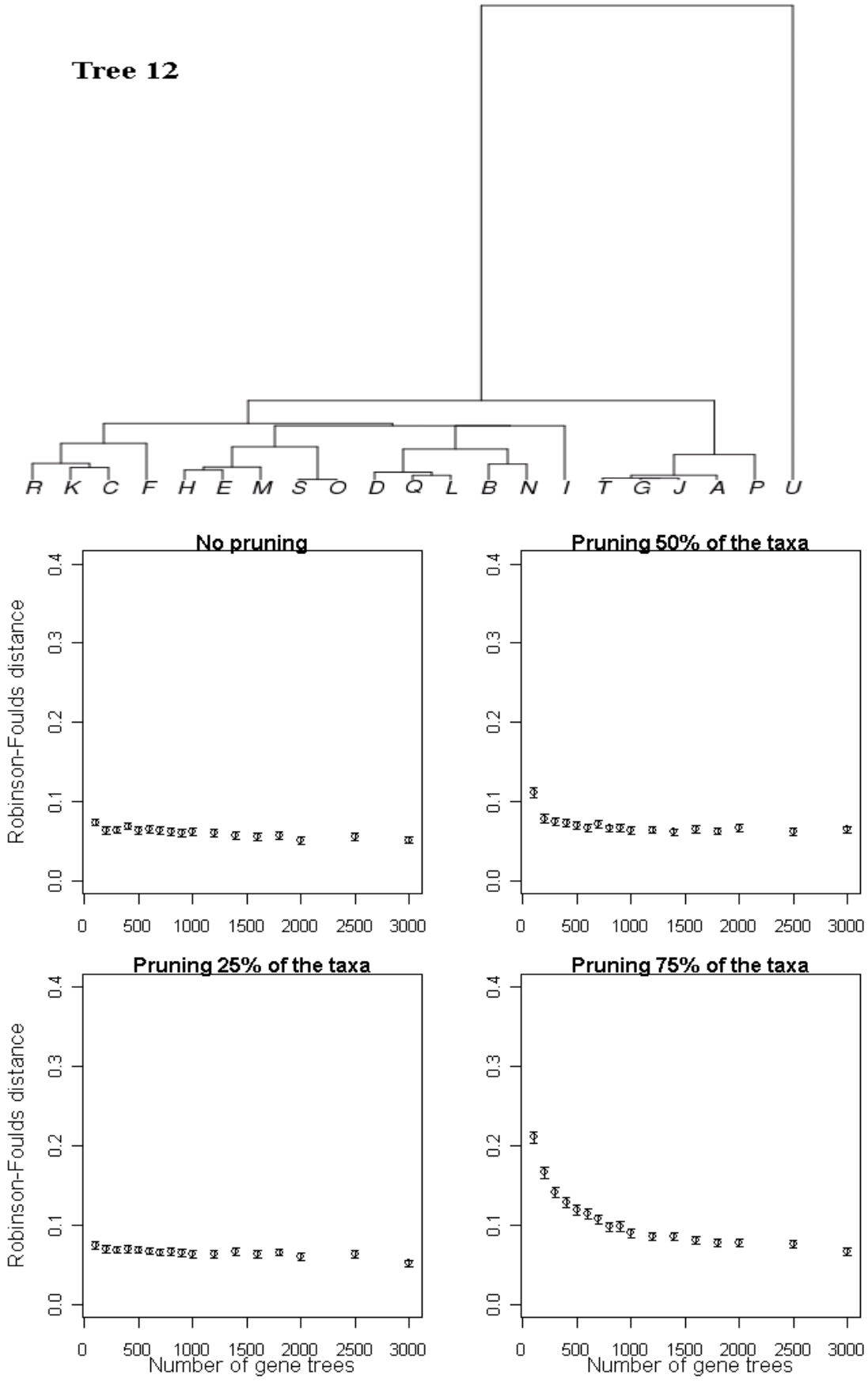


Figure 64. Tree 12 with simulated gene trees.

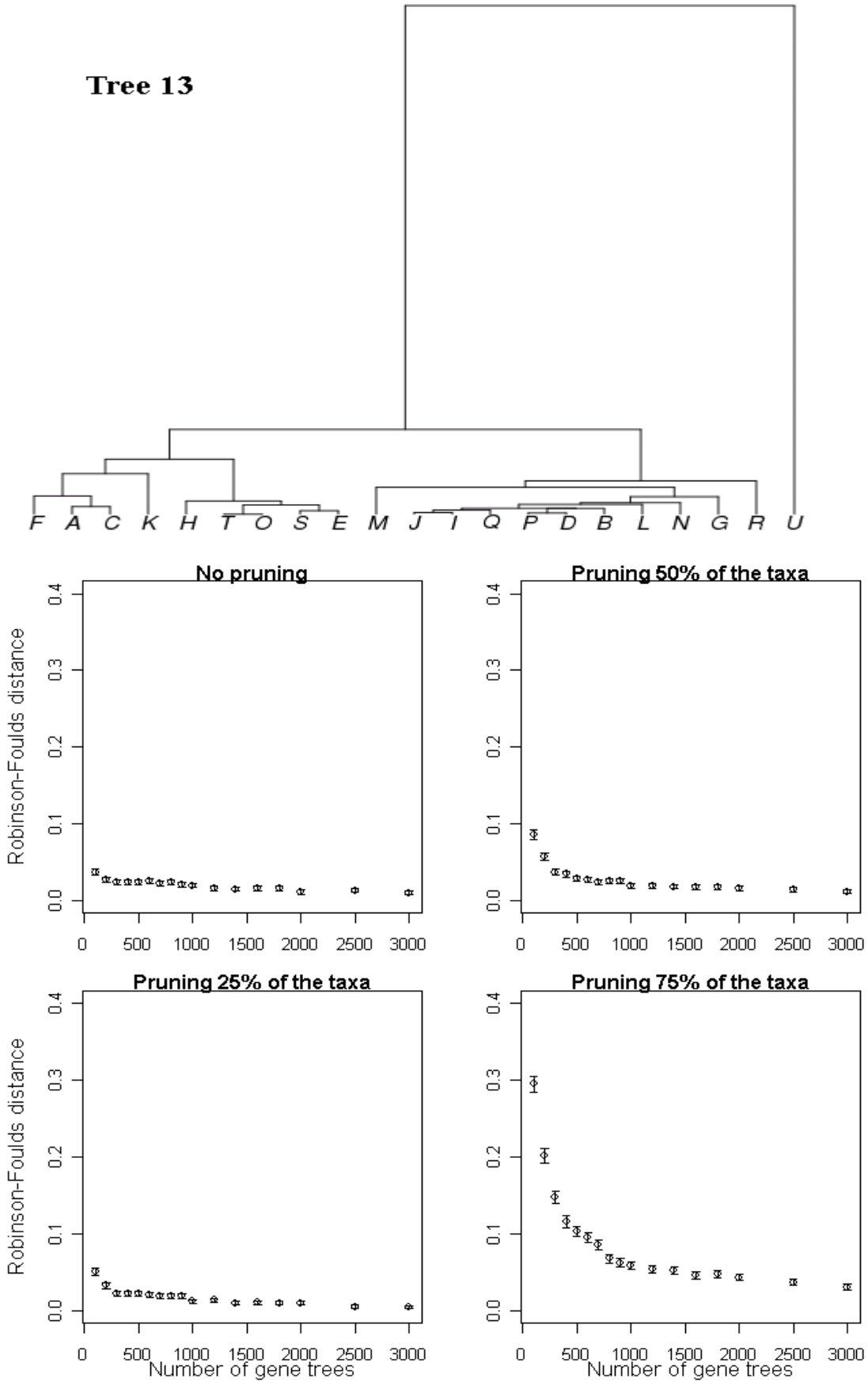


Figure 65. Tree 13 with simulated gene trees.

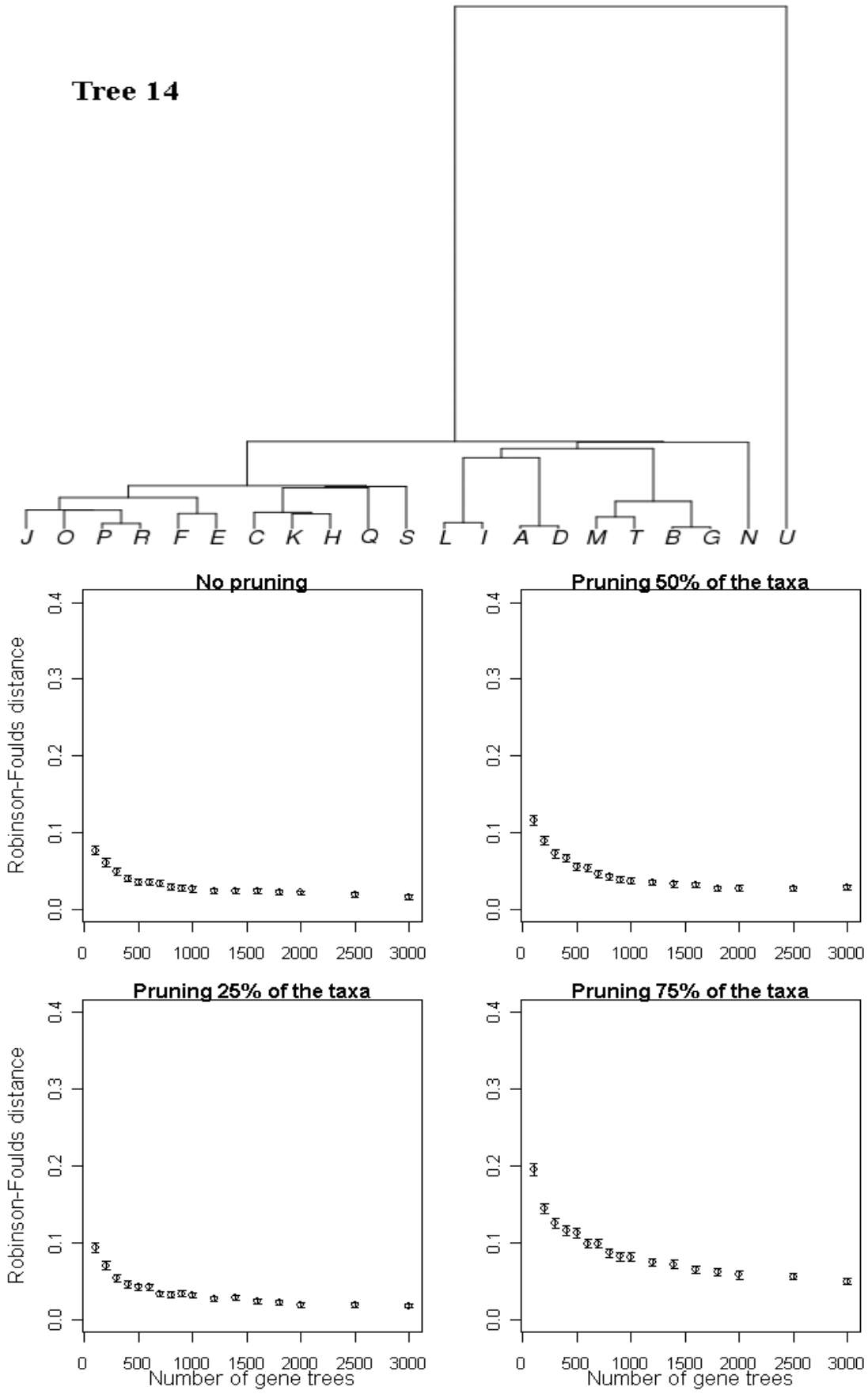


Figure 66. Tree 14 with simulated gene trees.

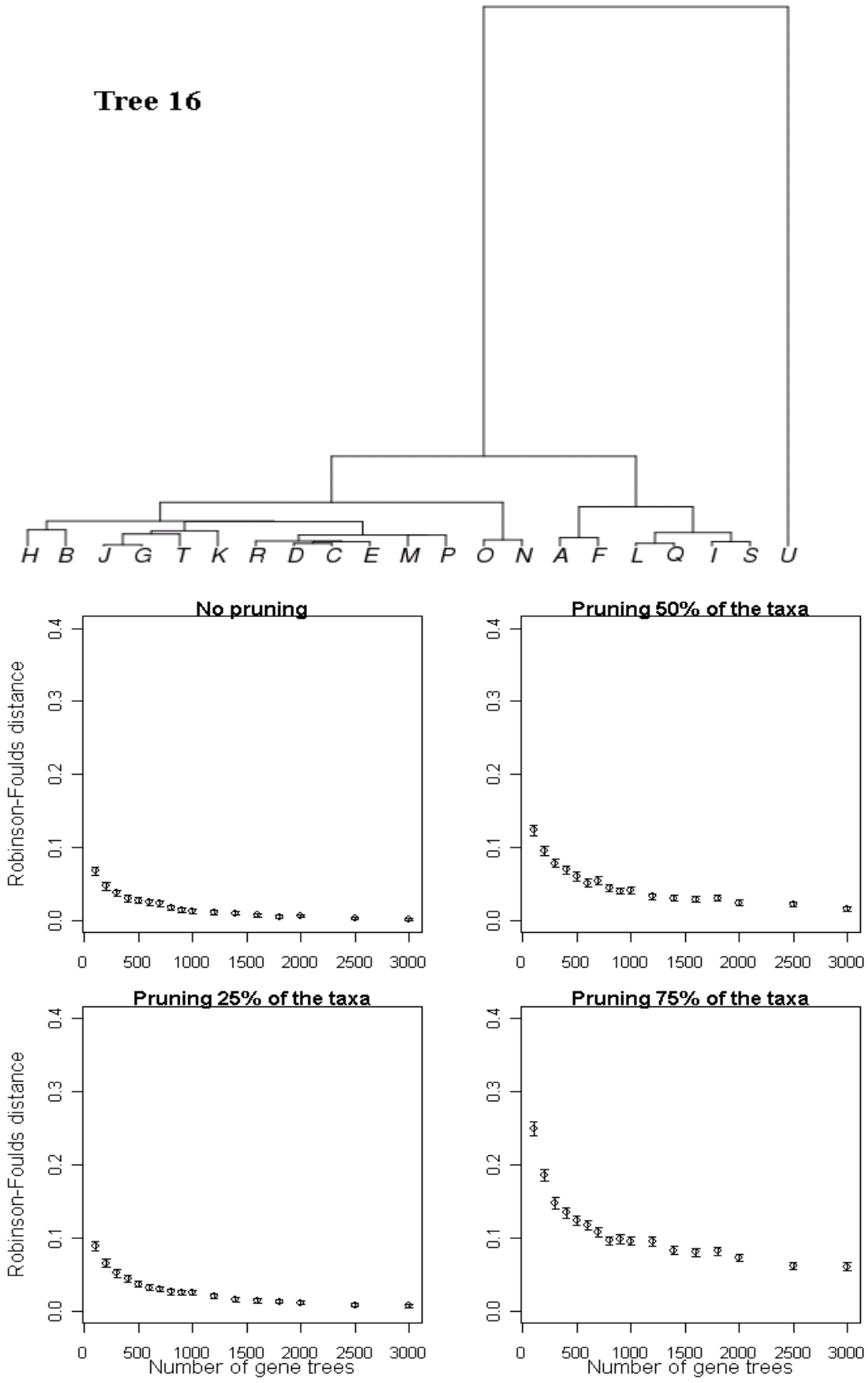


Figure 67. Tree 16 with simulated gene trees.

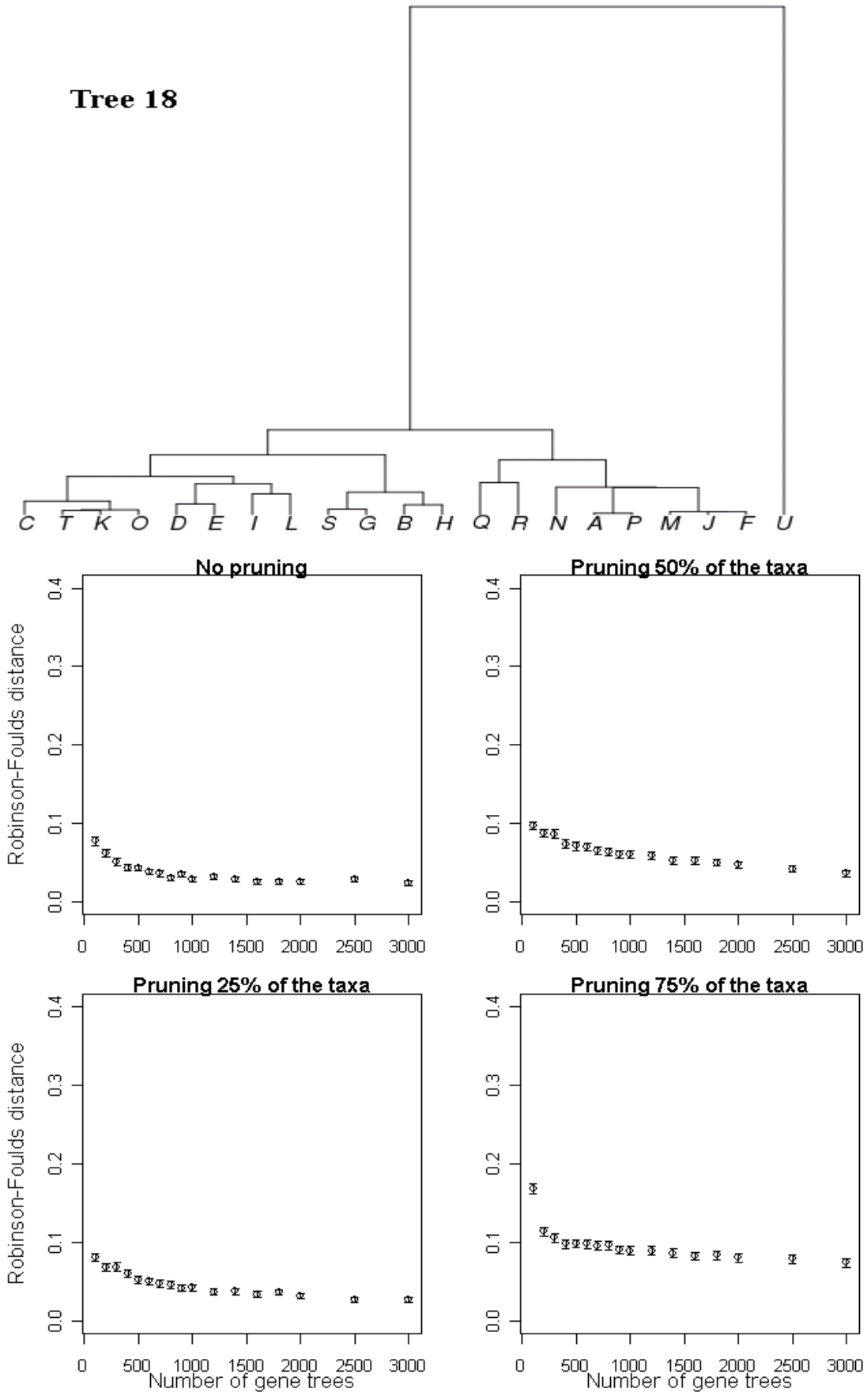


Figure 68. Tree 18 with simulated gene trees.

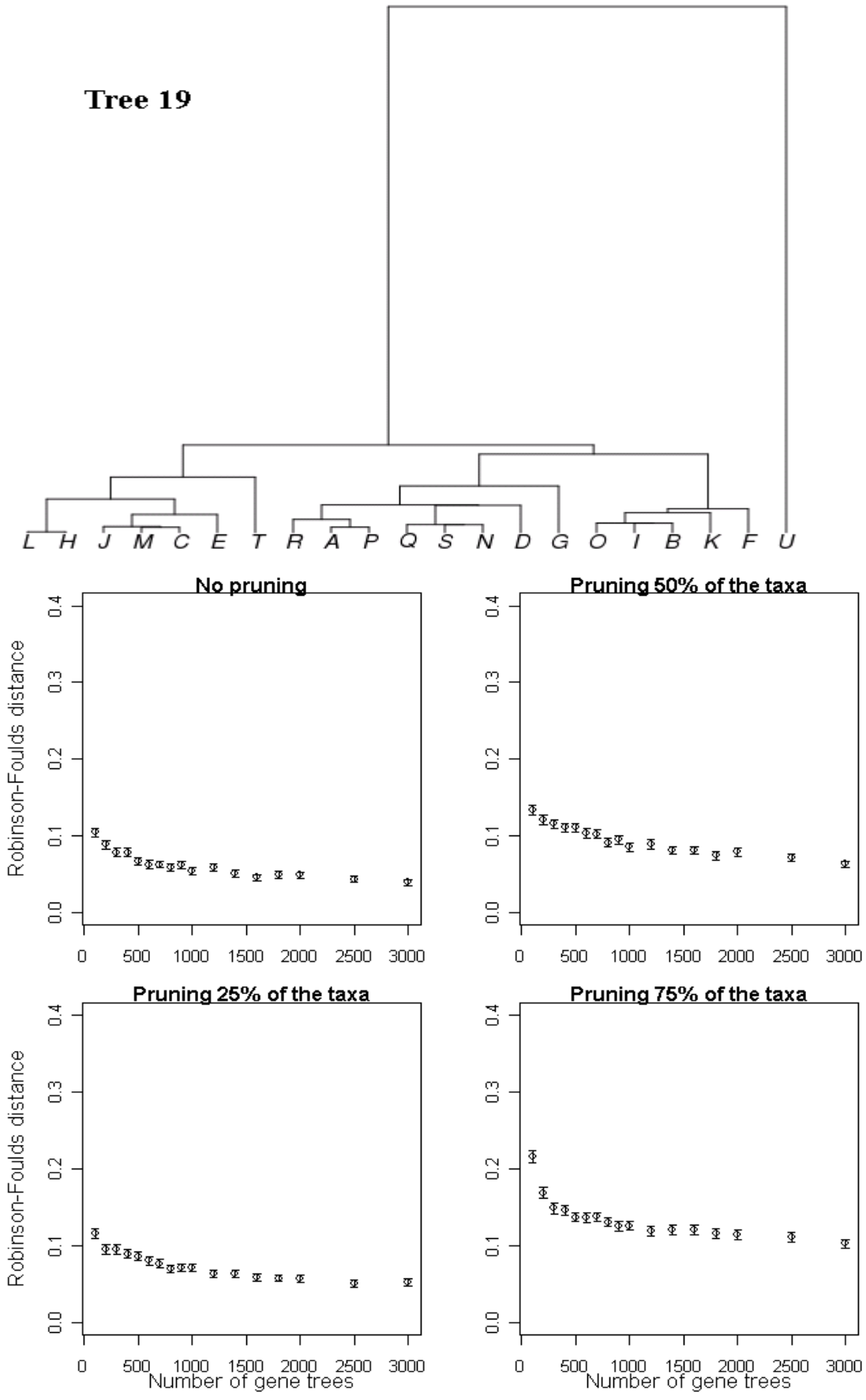


Figure 69. Tree 19 with simulated gene trees.

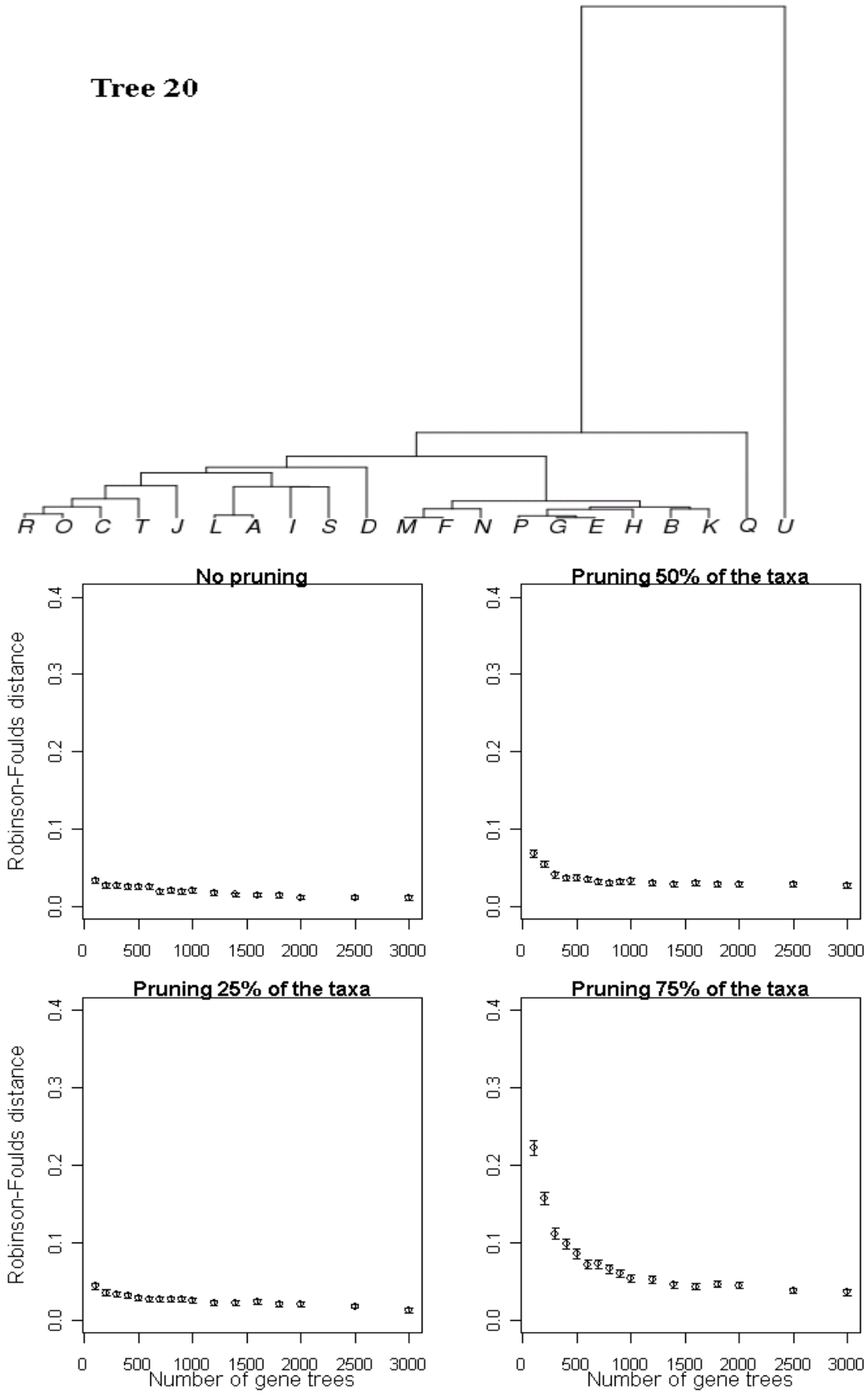


Figure 70. Tree 20 with simulated gene trees.

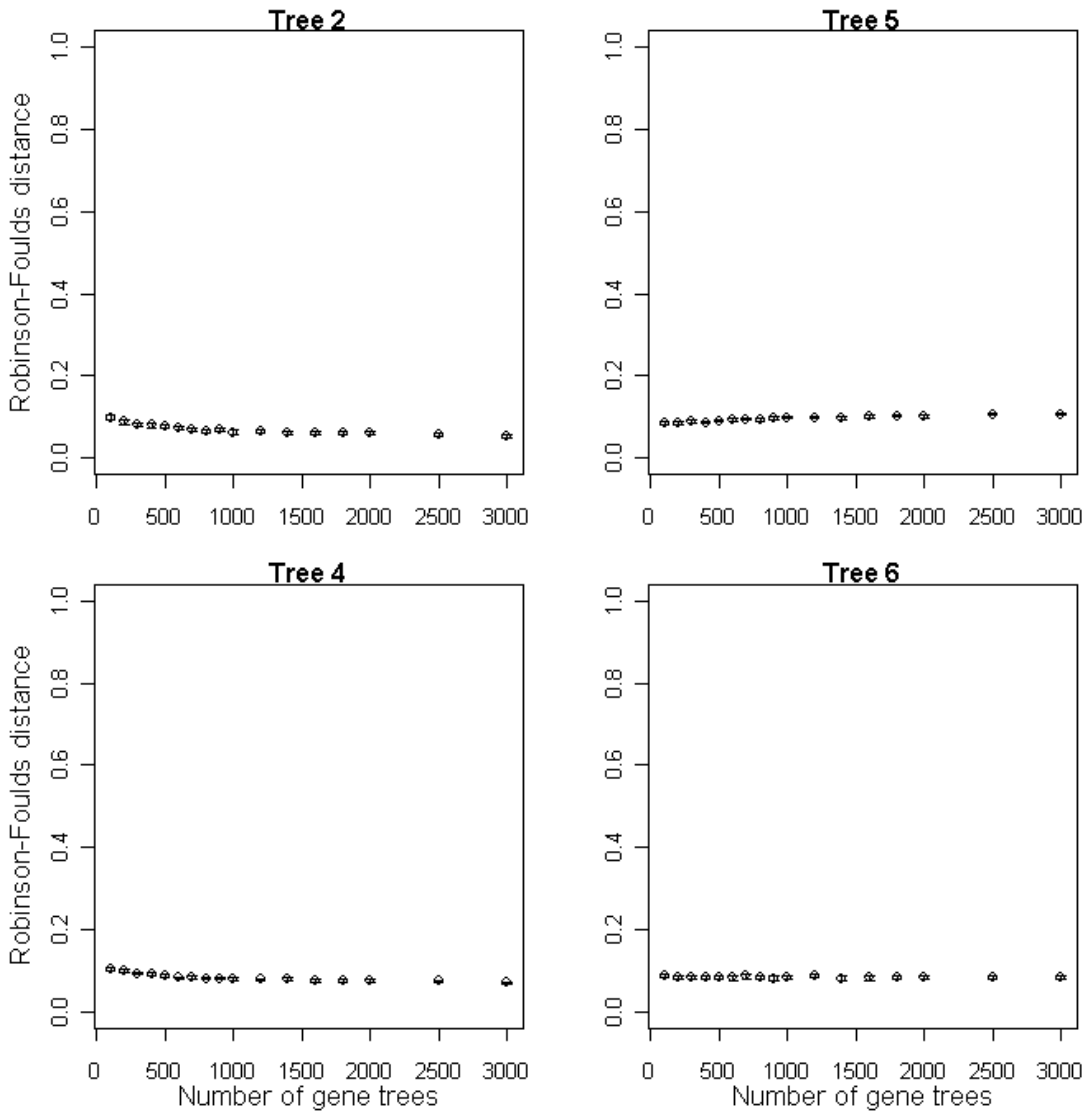


Figure 71. With estimated gene trees and no pruning.

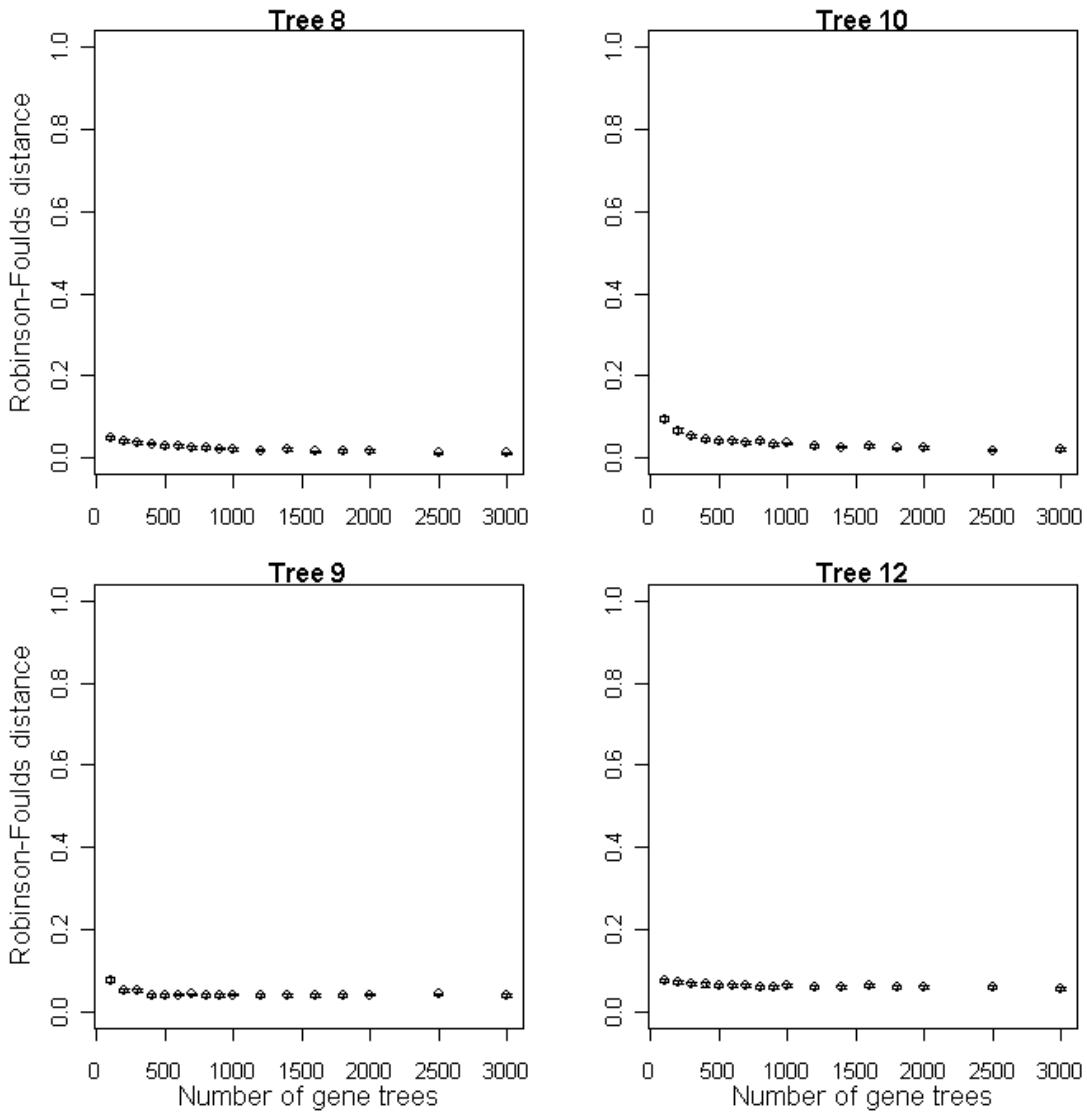


Figure 72. With estimated gene trees and no pruning.

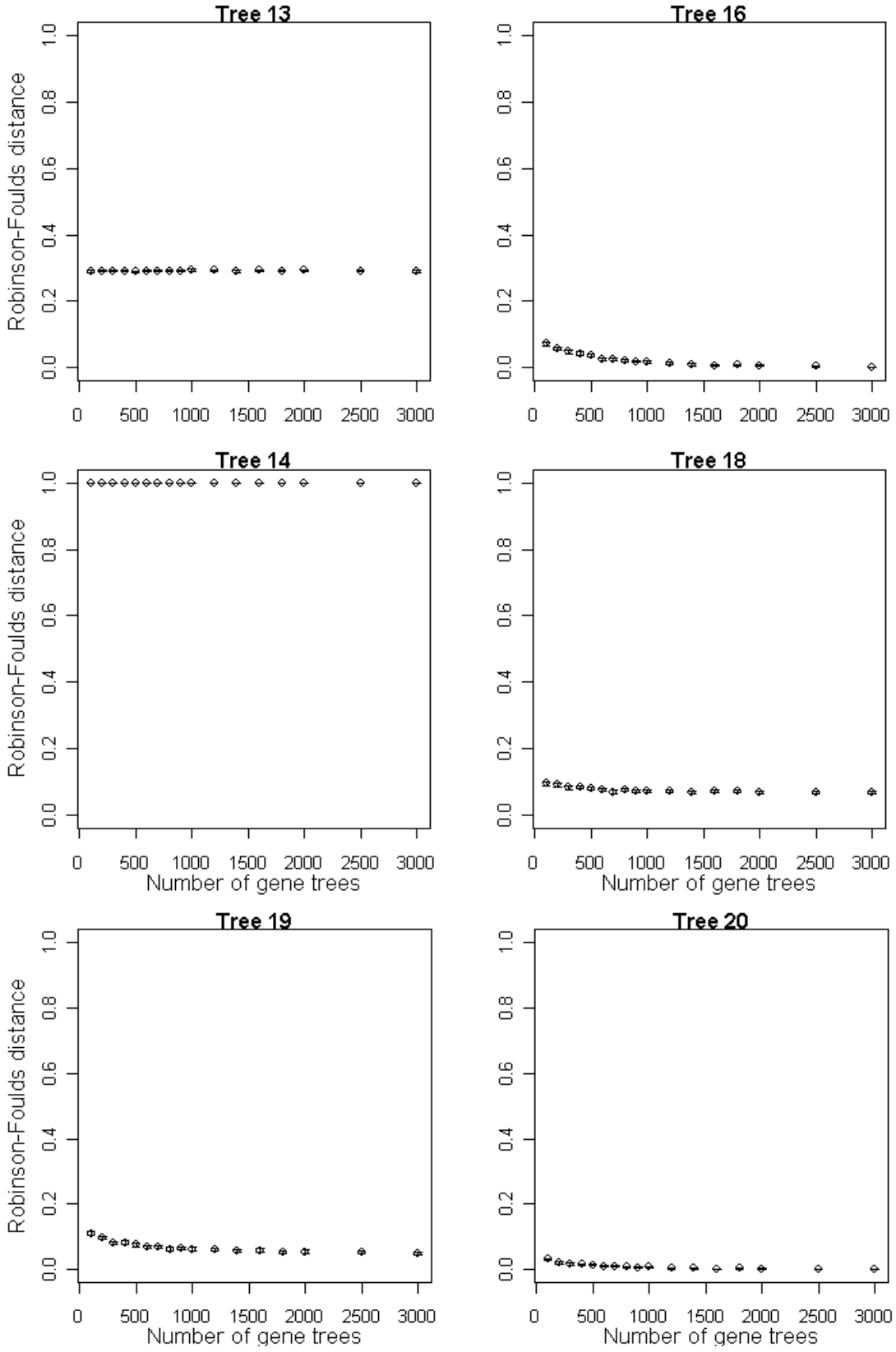


Figure 73. With estimated gene trees and no pruning.

Lists of full ordered 4-taxon and 5-taxon trees used to compute the expected value of parsimony score in Chapter 4. The outgroup is omitted.

For 4-taxon case:

tree t1 = (((A,B),C),D);
tree t2 = (((A,B),D),C);
tree t3 = (((A,C),B),D);
tree t4 = (((A,C),D),B);
tree t5 = (((A,D),B),C);
tree t6 = (((A,D),C),B);
tree t7 = (((B,C),A),D);
tree t8 = (((B,C),D),A);
tree t9 = (((B,D),A),C);
tree t10 = (((B,D),C),A);
tree t11 = (((C,D),A),B);
tree t12 = (((C,D),B),A);
tree t13 = ((A,B),(C,D));
tree t14 = ((A,C),(B,D));
tree t15 = ((A,D),(B,C));

For 5-taxon case:

tree t1 = ((((A,B),C),D),E);
tree t2 = ((((A,B),C),E),D);
tree t3 = ((((A,B),D),C),E);
tree t4 = ((((A,B),D),E),C);
tree t5 = ((((A,B),E),C),D);
tree t6 = ((((A,B),E),D),C);
tree t7 = ((((A,C),B),D),E);
tree t8 = ((((A,C),B),E),D);
tree t9 = ((((A,C),D),B),E);
tree t10 = ((((A,C),D),E),B);
tree t11 = ((((A,C),E),B),D);
tree t12 = ((((A,C),E),D),B);
tree t13 = ((((A,D),B),C),E);
tree t14 = ((((A,D),B),E),C);
tree t15 = ((((A,D),C),B),E);
tree t16 = ((((A,D),C),E),B);
tree t17 = ((((A,D),E),B),C);
tree t18 = ((((A,D),E),C),B);
tree t19 = ((((A,E),B),C),D);
tree t20 = ((((A,E),B),D),C);
tree t21 = ((((A,E),C),B),D);
tree t22 = ((((A,E),C),D),B);
tree t23 = ((((A,E),D),B),C);
tree t24 = ((((A,E),D),C),B);
tree t25 = ((((B,C),A),D),E);
tree t26 = ((((B,C),A),E),D);
tree t27 = ((((B,C),D),A),E);
tree t28 = ((((B,C),D),E),A);
tree t29 = ((((B,C),E),A),D);
tree t30 = ((((B,C),E),D),A);
tree t31 = ((((B,D),A),C),E);
tree t32 = ((((B,D),A),E),C);
tree t33 = ((((B,D),C),A),E);
tree t34 = ((((B,D),C),E),A);
tree t35 = ((((B,D),E),A),C);
tree t36 = ((((B,D),E),C),A);
tree t37 = ((((B,E),A),C),D);
tree t38 = ((((B,E),A),D),C);

tree t39 = ((((B,E),C),A),D);
tree t40 = ((((B,E),C),D),A);
tree t41 = ((((B,E),D),A),C);
tree t42 = ((((B,E),D),C),A);
tree t43 = ((((C,D),A),B),E);
tree t44 = ((((C,D),A),E),B);
tree t45 = ((((C,D),B),A),E);
tree t46 = ((((C,D),B),E),A);
tree t47 = ((((C,D),E),A),B);
tree t48 = ((((C,D),E),B),A);
tree t49 = ((((C,E),A),B),D);
tree t50 = ((((C,E),A),D),B);
tree t51 = ((((C,E),B),A),D);
tree t52 = ((((C,E),B),D),A);
tree t53 = ((((C,E),D),A),B);
tree t54 = ((((C,E),D),B),A);
tree t55 = ((((D,E),A),B),C);
tree t56 = ((((D,E),A),C),B);
tree t57 = ((((D,E),B),A),C);
tree t58 = ((((D,E),B),C),A);
tree t59 = ((((D,E),C),A),B);
tree t60 = ((((D,E),C),B),A);
tree t61 = (((A,B),(C,D)),E);
tree t62 = (((A,C),(B,D)),E);
tree t63 = (((A,D),(B,C)),E);
tree t64 = (((A,B),(C,E)),D);
tree t65 = (((A,C),(B,E)),D);
tree t66 = (((A,E),(B,C)),D);
tree t67 = (((A,B),(D,E)),C);
tree t68 = (((A,D),(B,E)),C);
tree t69 = (((A,E),(B,D)),C);
tree t70 = (((A,C),(D,E)),B);
tree t71 = (((A,D),(C,E)),B);
tree t72 = (((A,E),(C,D)),B);
tree t73 = (((B,C),(D,E)),A);
tree t74 = (((B,D),(C,E)),A);
tree t75 = (((B,E),(C,D)),A);
tree t76 = (((A,B),C),(D,E));
tree t77 = (((A,C),B),(D,E));
tree t78 = (((B,C),A),(D,E));
tree t79 = (((A,B),D),(C,E));
tree t80 = (((A,D),B),(C,E));

tree t81 = (((B,D),A),(C,E));
tree t82 = (((A,C),D),(B,E));
tree t83 = (((A,D),C),(B,E));
tree t84 = (((C,D),A),(B,E));
tree t85 = (((B,C),D),(A,E));
tree t86 = (((B,D),C),(A,E));
tree t87 = (((C,D),B),(A,E));
tree t88 = (((A,B),E),(C,D));
tree t89 = (((A,E),B),(C,D));
tree t90 = (((B,E),A),(C,D));
tree t91 = (((A,C),E),(B,D));
tree t92 = (((A,E),C),(B,D));
tree t93 = (((C,E),A),(B,D));
tree t94 = (((B,C),E),(A,D));
tree t95 = (((B,E),C),(A,D));
tree t96 = (((C,E),B),(A,D));
tree t97 = (((A,D),E),(B,C));
tree t98 = (((A,E),D),(B,C));
tree t99 = (((D,E),A),(B,C));
tree t100 = (((B,D),E),(A,C));
tree t101 = (((B,E),D),(A,C));
tree t102 = (((D,E),B),(A,C));
tree t103 = (((C,D),E),(A,B));
tree t104 = (((C,E),D),(A,B));
tree t105 = (((D,E),C),(A,B));

Construct the EPS_T equation of a candidate tree from the ordered 4-taxon list.

The consensus part EPS_C:

	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15	
A	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	0	1	0	0	0	1	0	0	1	0	1	0	1	0
B	1	1	1	1	0	1	0	0	0	1	0	0	1	1	1	1	1	1	1	1	0	0	0	1	1	0	0	1	0	1
C	0	1	0	0	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	1	0	1	1	0	0	1
D	0	0	0	1	0	0	0	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0

Figure 74. The matrix representation, each tree coded into 2 columns of 0's and 1's. 0 indicates that particular taxon does not occur, and 1 if occur.

	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15	
PS against t1	1	1	1	2	2	1	2	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2
PS against t2	1	2	1	1	2	2	2	2	2	1	2	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2	2	2	2	2
PS against t3	2	1	2	2	1	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2
PS against t4	2	2	2	2	1	2	1	1	2	2	2	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	1	2	2	2
PS against t5	2	2	2	1	2	2	2	2	1	1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	1	2
PS against t6	2	2	2	2	2	2	2	1	1	2	1	1	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2	1	2
PS against t7	2	1	2	2	2	1	2	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2
PS against t8	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	1	2	2	2	1	2	2	2	1	2	2	2	2	1	2
PS against t9	2	2	2	1	2	2	2	2	2	1	2	2	2	2	2	2	1	1	1	2	2	2	2	2	2	2	1	2	2	2
PS against t10	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2	1	1	2	2	2	1	2	2	1	2	2	2
PS against t11	2	2	2	2	2	2	2	1	2	2	2	1	2	2	2	2	2	2	2	2	1	1	1	2	2	1	2	2	2	2
PS against t12	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	1	1	2	1	1	1	2	2	2	2	2
PS against t13	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	1	1	1	2	2	2	2
PS against t14	2	2	2	2	1	2	1	2	2	2	2	2	2	2	2	2	2	1	2	1	2	2	2	2	2	2	1	1	2	2
PS against t15	2	2	2	2	2	2	2	2	1	2	1	2	2	1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1	1

Figure 75.The corresponding parsimony scores (PS) against each tree in the consensus setting.

The supertree part EPS_S:

	1	2	3	4	5	6	7	8
A	? 1 1 1	? 1 1 1	? 1 1 1	? 1 1 1	? 1 1 1	? 1 1 1	? 1 1 0	? 0 0 0
B	1 ? 1 1	1 ? 1 1	1 ? 1 0	0 ? 0 0	1 ? 0 1	0 ? 0 0	1 ? 1 1	1 ? 1 1
C	1 1 ? 0	0 0 ? 0	1 1 ? 1	1 1 ? 1	0 0 ? 0	1 0 ? 1	1 1 ? 1	1 1 ? 1
D	0 0 0 ?	1 1 0 ?	0 0 0 ?	1 0 1 ?	1 1 1 ?	1 1 1 ?	0 0 0 ?	0 1 1 ?

	9	10	11	12	13	14	15
A	? 1 0 1	? 0 0 0	? 0 1 1	? 0 0 0	? 0 1 1	? 1 0 1	? 1 1 0
B	1 ? 1 1	1 ? 1 1	0 ? 0 0	0 ? 1 1	0 ? 1 1	1 ? 1 0	1 ? 0 1
C	0 0 ? 0	0 1 ? 1	1 1 ? 1	1 1 ? 1	1 1 ? 0	0 1 ? 1	1 0 ? 1
D	1 1 1 ?	1 1 1 ?	1 1 1 ?	1 1 1 ?	1 1 0 ?	1 0 1 ?	0 1 1 ?

Figure 76. The matrix representation,
each tree coded into 4 columns of 0's, 1's and ?'s.
The first column is where taxon A is deleted,
(or taxon A is a missing data, indicated as the ? mark) and etc.

	1	2	3	4	5	6	7	8
PS against t1	1 1 1 1	2 2 1 1	1 1 1 2	2 1 2 2	2 2 2 1	2 2 2 2	1 1 1 2	1 2 2 2
PS against t2	2 2 1 1	1 1 1 1	2 2 1 2	2 2 2 2	1 1 2 1	2 1 2 2	2 2 1 2	2 2 2 2
PS against t3	1 1 1 2	2 2 1 2	1 1 1 1	2 1 2 1	2 2 2 2	2 2 2 1	1 1 1 2	1 2 2 2
PS against t4	2 1 2 2	2 2 2 2	2 1 2 1	1 1 1 1	2 2 1 2	1 2 1 1	2 1 2 2	2 2 2 2
PS against t5	2 2 2 1	1 1 2 1	2 2 2 2	2 2 1 2	1 1 1 1	2 1 1 2	2 2 2 2	2 2 2 2
PS against t6	2 2 2 2	2 1 2 2	2 2 2 1	1 2 1 1	2 1 1 2	1 1 1 1	2 2 2 2	2 2 2 2
PS against t7	1 1 1 2	2 2 1 2	1 1 1 2	2 1 2 2	2 2 2 2	2 2 2 2	1 1 1 1	1 2 2 1
PS against t8	1 2 2 2	2 2 2 2	1 2 2 2	2 2 2 2	2 2 2 2	2 2 2 2	1 2 2 1	1 1 1 1
PS against t9	2 2 2 1	1 1 2 1	2 2 2 2	2 2 2 2	1 1 2 1	2 1 2 2	2 2 2 2	2 2 1 2
PS against t10	2 2 2 2	1 2 2 2	2 2 2 2	2 2 2 2	1 2 2 2	2 2 2 2	2 2 2 1	2 1 1 1
PS against t11	2 2 2 2	2 2 2 2	2 2 2 1	1 2 1 1	2 2 1 2	1 2 1 1	2 2 2 2	2 1 2 2
PS against t12	2 2 2 2	2 2 2 2	2 2 2 2	1 2 2 2	2 2 2 2	1 2 2 2	2 2 2 1	2 1 1 1
PS against t13	2 2 1 1	2 2 1 1	2 2 1 2	1 2 2 2	2 2 2 1	1 2 2 2	2 2 1 2	2 1 2 2
PS against t14	2 1 2 2	1 2 2 2	2 1 2 1	2 1 2 1	1 2 2 2	2 2 2 1	2 1 2 2	2 2 1 2
PS against t15	1 2 2 2	2 1 2 2	1 2 2 2	2 2 1 2	2 1 1 2	2 1 1 2	1 2 2 1	1 2 2 1

	9	10	11	12	13	14	15
PS against t1	2 2 2 1	2 2 2 2	2 2 2 2	2 2 2 2	2 2 1 1	2 1 2 2	1 2 2 2
PS against t2	1 1 2 1	1 2 2 2	2 2 2 2	2 2 2 2	2 2 1 1	1 2 2 2	2 1 2 2
PS against t3	2 2 2 2	2 2 2 2	2 2 2 1	2 2 2 2	2 2 1 2	2 1 2 1	1 2 2 2
PS against t4	2 2 2 2	2 2 2 2	1 2 1 1	1 2 2 2	1 2 2 2	2 1 2 1	2 2 1 2
PS against t5	1 1 2 1	1 2 2 2	2 2 1 2	2 2 2 2	2 2 2 1	1 2 2 2	2 1 1 2
PS against t6	2 1 2 2	2 2 2 2	1 2 1 1	1 2 2 2	1 2 2 2	2 2 2 1	2 1 1 2
PS against t7	2 2 2 2	2 2 2 1	2 2 2 2	2 2 2 1	2 2 1 2	2 1 2 2	1 2 2 1
PS against t8	2 2 1 2	2 1 1 1	2 1 2 2	2 1 1 1	2 1 2 2	2 2 1 2	1 2 2 1
PS against t9	1 1 1 1	1 2 1 2	2 2 2 2	2 2 1 2	2 2 2 1	1 2 1 2	2 1 2 2
PS against t10	1 2 1 2	1 1 1 1	2 1 2 2	2 1 1 1	2 1 2 2	1 2 1 2	2 2 2 1
PS against t11	2 2 2 2	2 1 2 2	1 1 1 1	1 1 2 2	1 1 2 2	2 2 2 1	2 2 1 2
PS against t12	2 2 1 2	2 1 1 1	1 1 2 2	1 1 1 1	1 1 2 2	2 2 1 2	2 2 2 1
PS against t13	2 2 2 1	2 1 2 2	1 1 2 2	1 1 2 2	1 1 1 1	2 2 2 2	2 2 2 2
PS against t14	1 2 1 2	1 2 1 2	2 2 2 1	2 2 1 2	2 2 2 2	1 1 1 1	2 2 2 2
PS against t15	2 1 2 2	2 2 2 1	2 2 1 2	2 2 2 1	2 2 2 2	2 2 2 2	1 1 1 1

Figure 77.The corresponding parsimony scores (PS) against each tree in the supertree setting. The first column of each tree represents the case where taxon A is deleted and so forth.

Then, the goal is to find the EPS_C and EPS_S by weighting corresponding parsimony score with the probability of occurring of each gene tree, $p_1, p_2, \cdots, p_{15}$ that sum up to 1 exactly.

For example, for tree 1 (the second rows of Figures 75 and 77), we have:

$$EPS_{C1} = (1+1)p_1 + (1+2)p_2 + (2+1)p_3 + (2+2)p_4 + (2+2)p_5 \\ +(2+2)p_6 + (2+1)p_7 + (2+2)p_8 + (2+2)p_9 + (2+2)p_{10} \\ +(2+2)p_{11} + (2+2)p_{12} + (1+2)p_{13} + (2+2)p_{14} + (2+2)p_{15},$$

and similarly

$$EPS_{S1} = \\ d_A\{(p_1 + p_3 + p_7 + p_8 + p_{15}) \times 1 \\ + (p_2 + p_4 + p_5 + p_6 + p_9 + p_{10} + p_{11} + p_{12} + p_{13} + p_{14}) \times 2\} \\ +d_B\{(p_1 + p_3 + p_4 + p_7 + p_{14}) \times 1 \\ + (p_2 + p_5 + p_6 + p_8 + p_9 + p_{10} + p_{11} + p_{12} + p_{13} + p_{15}) \times 2\} \\ +d_C\{(p_1 + p_2 + p_3 + p_7 + p_{13}) \times 1 \\ + (p_4 + p_5 + p_6 + p_8 + p_9 + p_{10} + p_{11} + p_{12} + p_{14} + p_{15}) \times 2\} \\ +d_D\{(p_1 + p_2 + p_5 + p_9 + p_{13}) \times 1 \\ + (p_3 + p_4 + p_6 + p_7 + p_8 + p_{10} + p_{11} + p_{12} + p_{14} + p_{15}) \times 2\}.$$

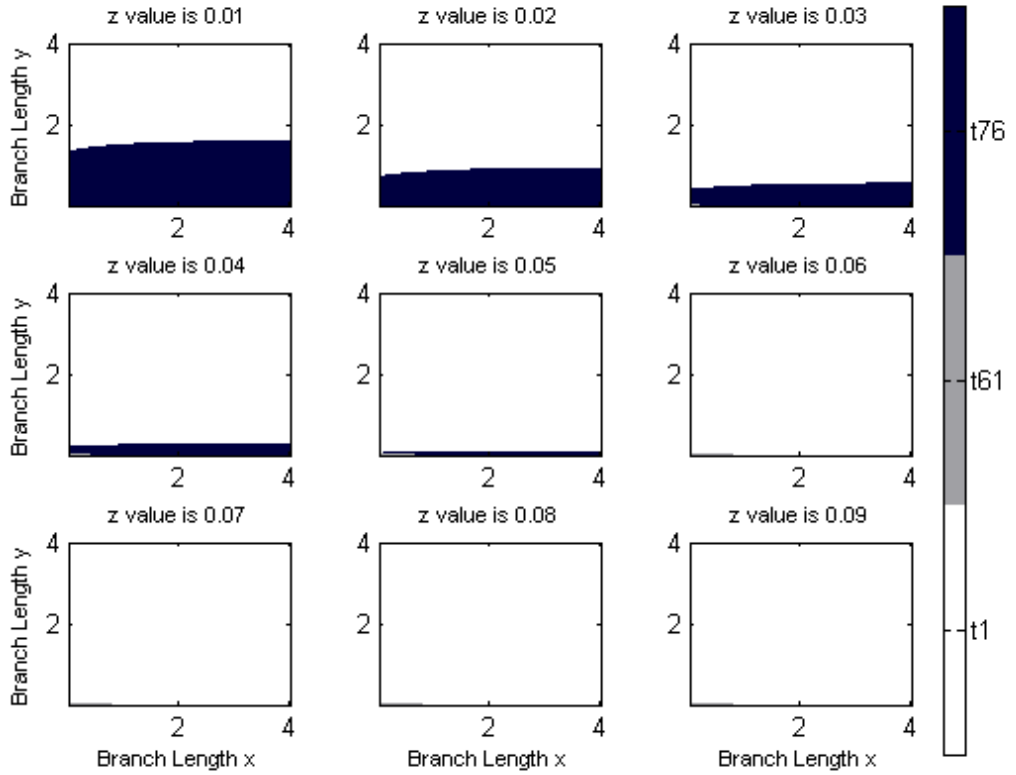
Finally, one only need to weight the EPS_C and EPS_S by the pruning scheme weight factor w to produce EPS_t. For this tree 1 example $EPS_{t1} = (1-w) \times EPS_{C1} + w \times EPS_{S1}$. Similarly, one can follow this procedure to obtain the EPS_t equations for the rest trees in the 4-taxon trees list, and also for the 5-taxon trees list.

Extra results of the analytical equation for 5-taxon species trees

Recall that ((((A,B),C),D),E), (((A,B),(C,D)),E) and (((A,B),C),(D,E)) are the 3 species tree topologies in Section 4.3 with the same setting. Then, for the true species tree topology ((((A,B),C),D),E), only *t1*, *t61* and *t76* are considered. Under the true species tree topology (((A,B),(C,D)),E), and the EPS_t are only calculated for 3 candidate trees, *t67*, *t88* and *t103*. Similarly, if the true species tree topology is (((A,B),C),(D,E)), and the EPS_t are only calculated for *t76* and *t105*. Where, *t1* = ((((A,B),C),D),E), *t61* = (((A,B),(C,D)),E), *t76* = (((A,B),C),(D,E)), *t88* = (((A,B),E),(C,D)), *t103* = ((A,B),(E,(C,D))) and *t105* = ((A,B),(C,(D,E))).

Let D_1 and D_2 indicate that exactly 1 or 2 taxa are randomly deleted from the gene trees, and the probability of deleting each taxon A, B, C, D and E is equal so that $d_A = d_B = d_C = d_D = d_E = ⅕$. On top of that, let *C* and *S* denote the consensus setting, no pruning and supertree setting, at least 1taxon randomly pruned, respectively. Then, the following 4 pruning schemes are used: (i) $½C + ½S\ (1D_1 + 0D_2)$; (ii) $½C + ½S\ (½D_1 + ½D_2)$; (iii) $½C + ½S\ (0D_1 + 1D_2)$ and (iv) $0C + 1S\ (½D_1 + ½D_2)$ to produce some plots to indicate the MRPAST under different true species tree topologies. Again, for the branch lengths (*x*, *y*, *z*), *z* is fixed and varying *x* and *y*.

(a)



(b)

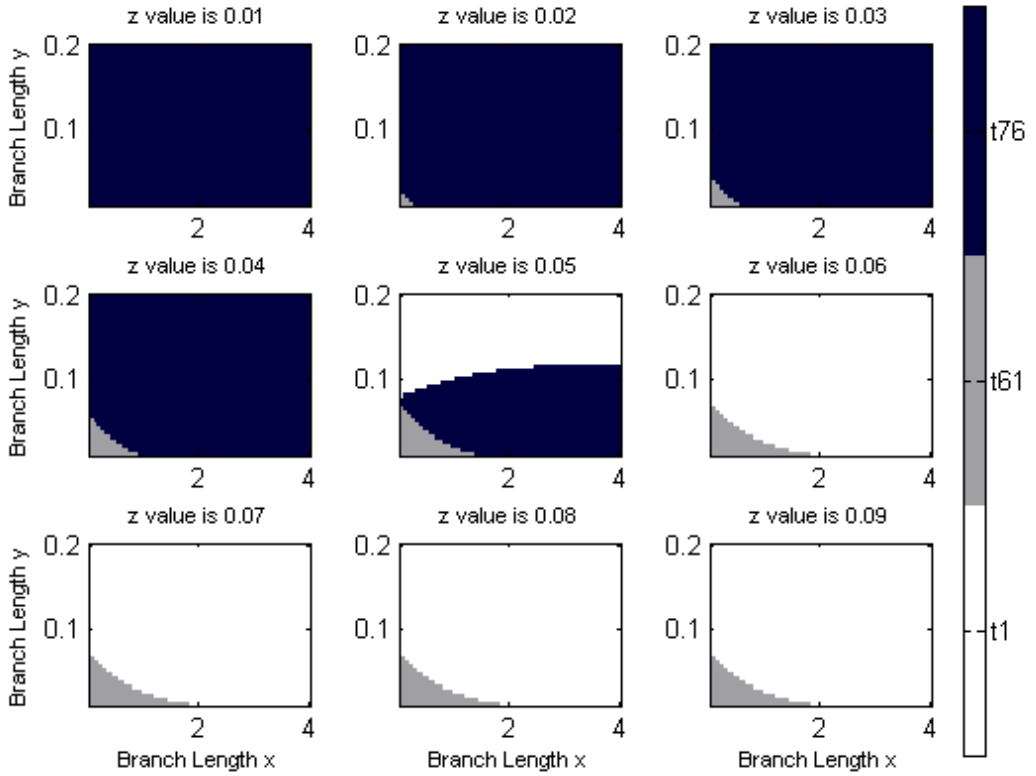


Figure 78. With pruning scheme ½*C* + ½*S* (1D_1+ 0D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

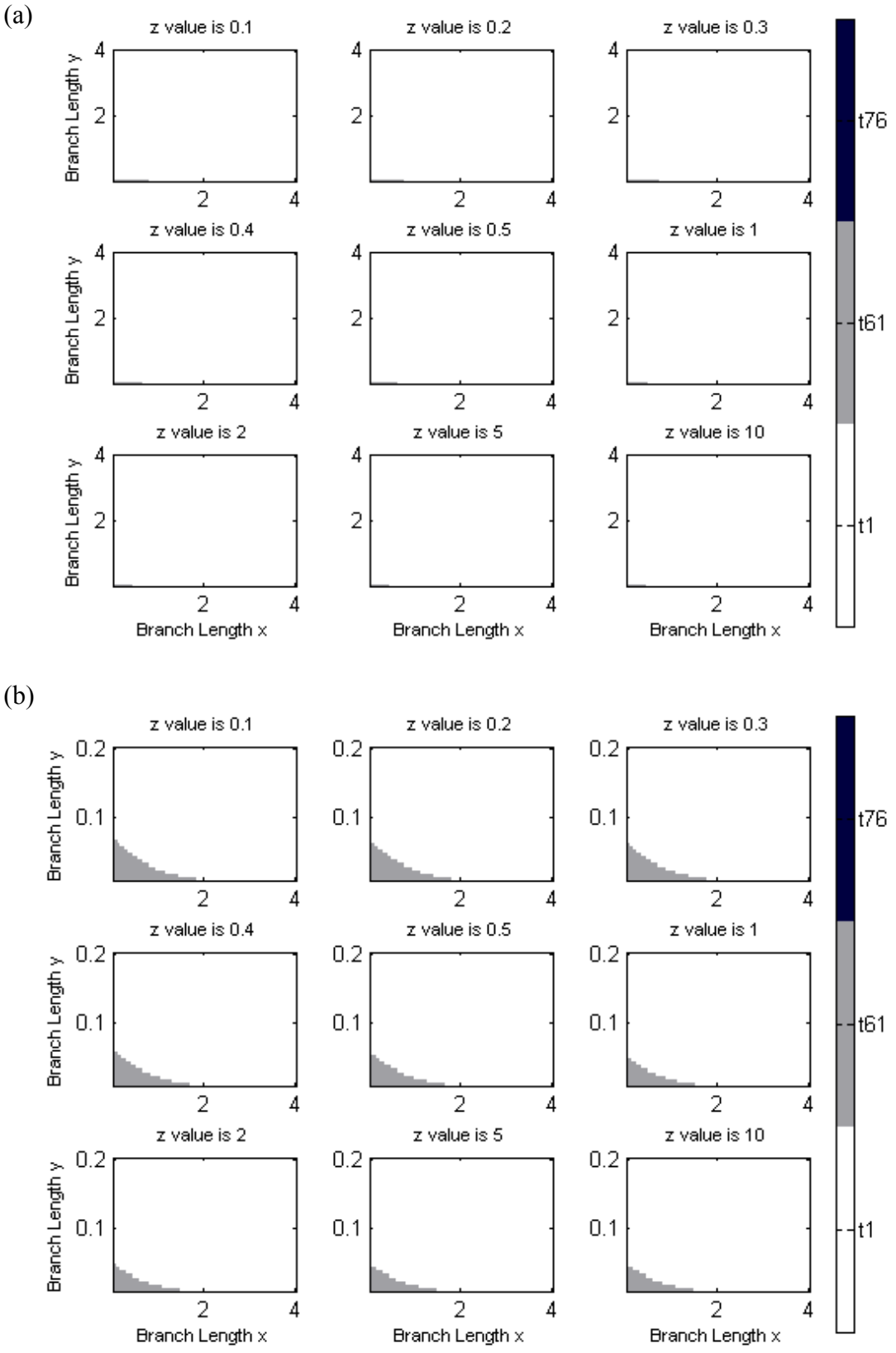


Figure 79. With pruning scheme ½*C* + ½*S* (1D_1+ 0D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

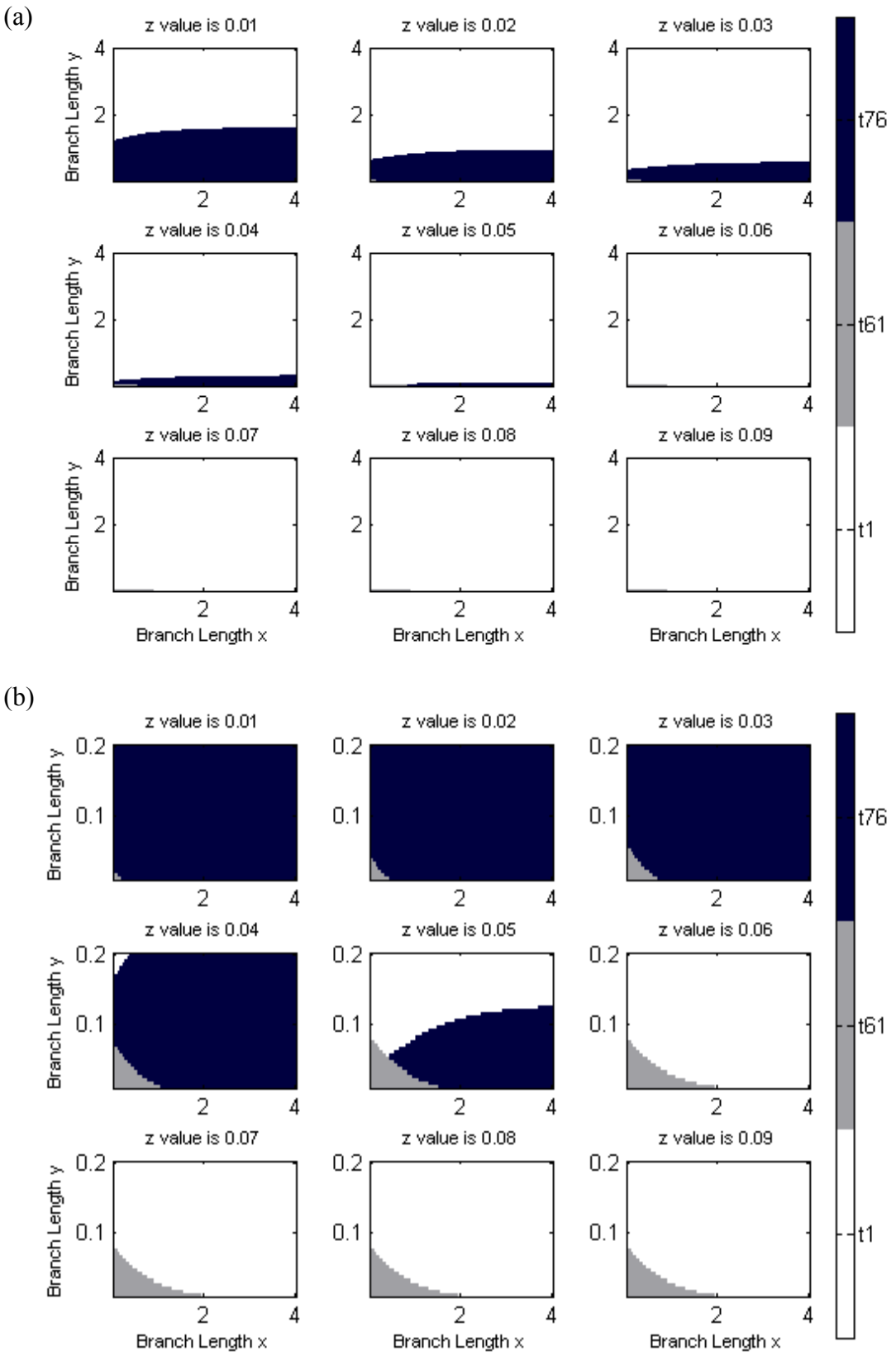


Figure 80. With pruning scheme ½*C* + ½*S* (½D_1+ ½D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

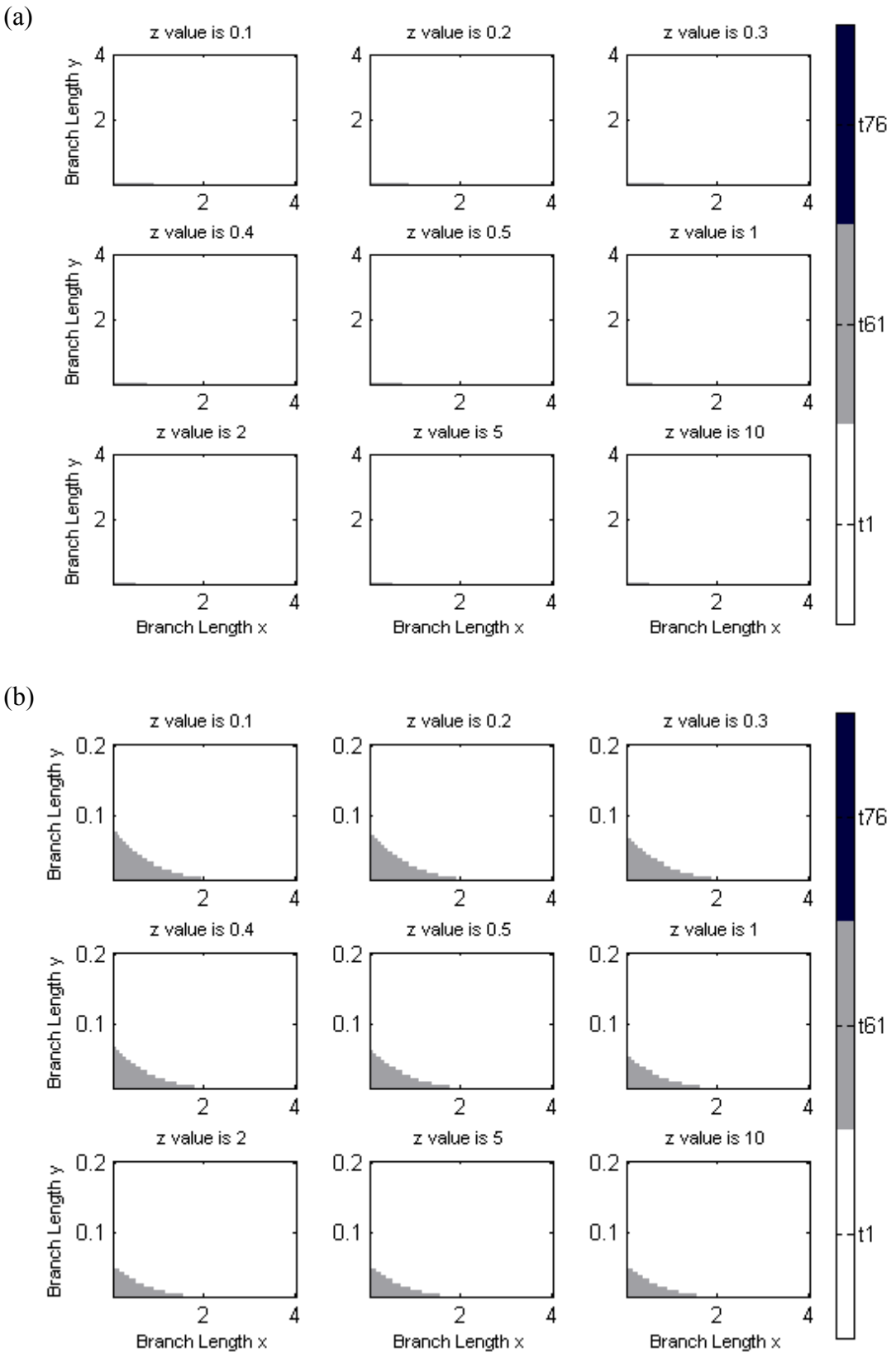


Figure 81. With pruning scheme ½*C* + ½*S* (½D_1+ ½D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

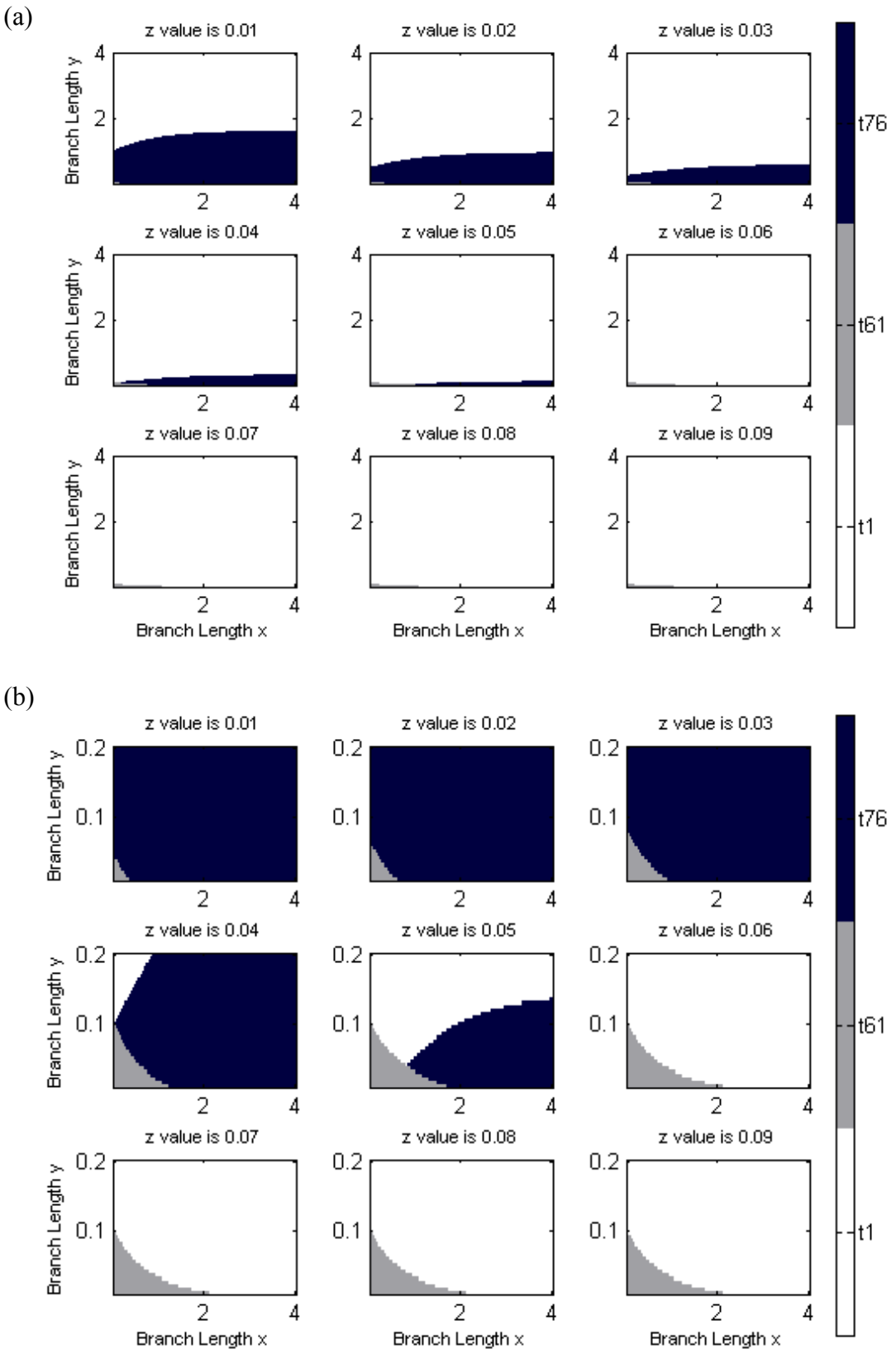


Figure 82. With pruning scheme ½*C* + ½*S* (0D_1+ 1D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

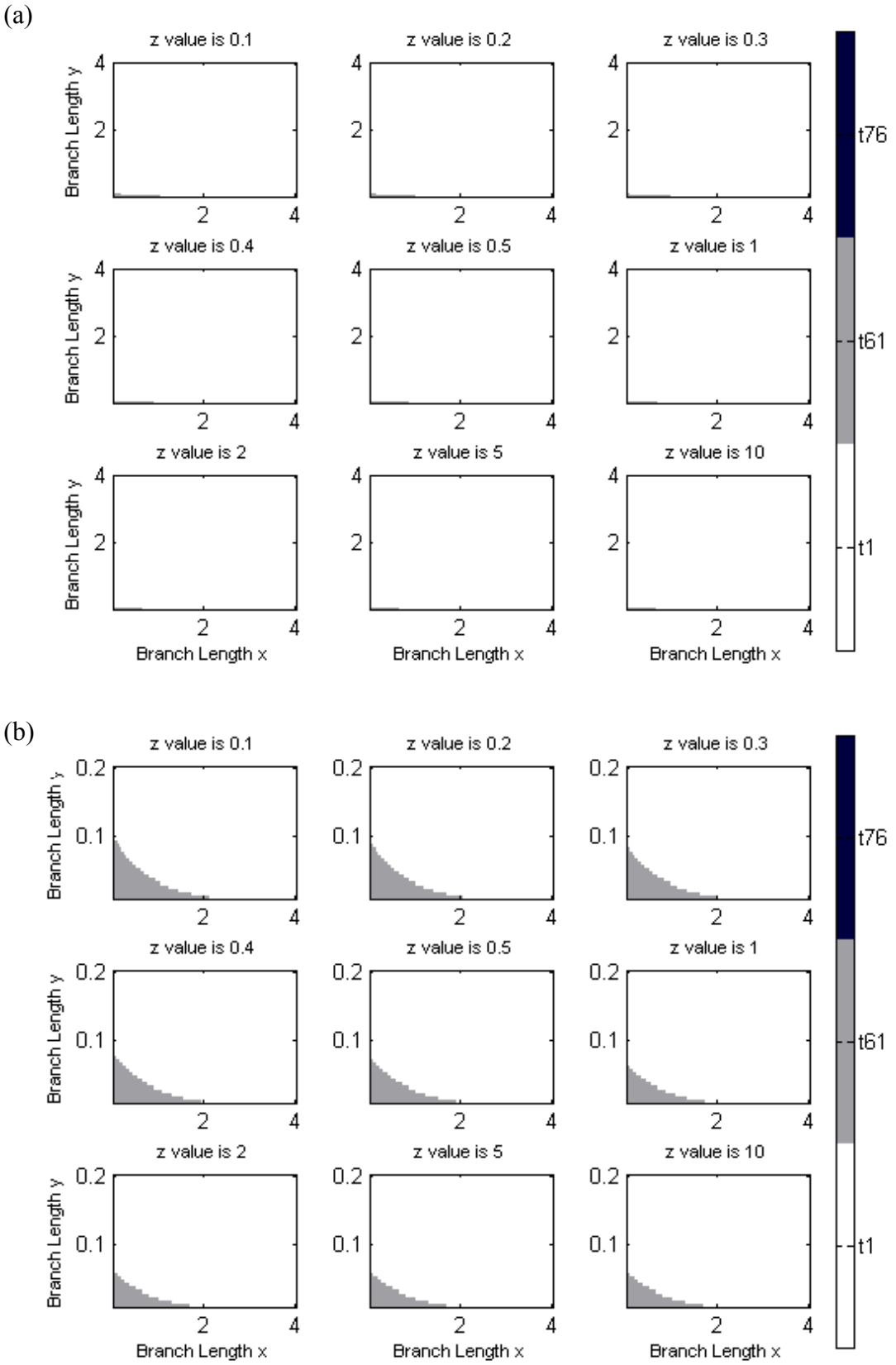


Figure 83. With pruning scheme ½*C* + ½*S* (0D_1+ 1D_2) under the true species tree topology t1. Part (b) is a zoom in of part (a).

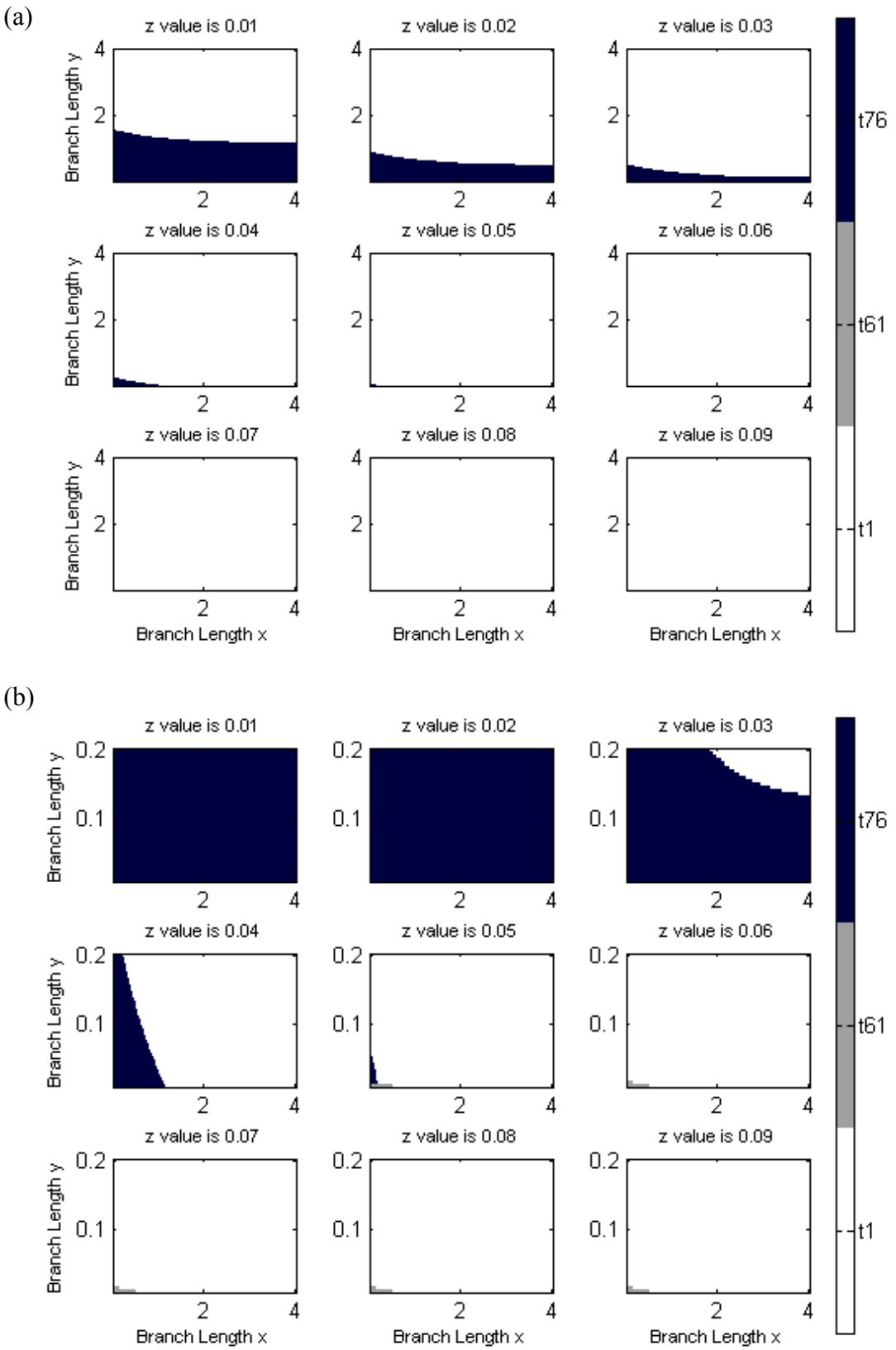
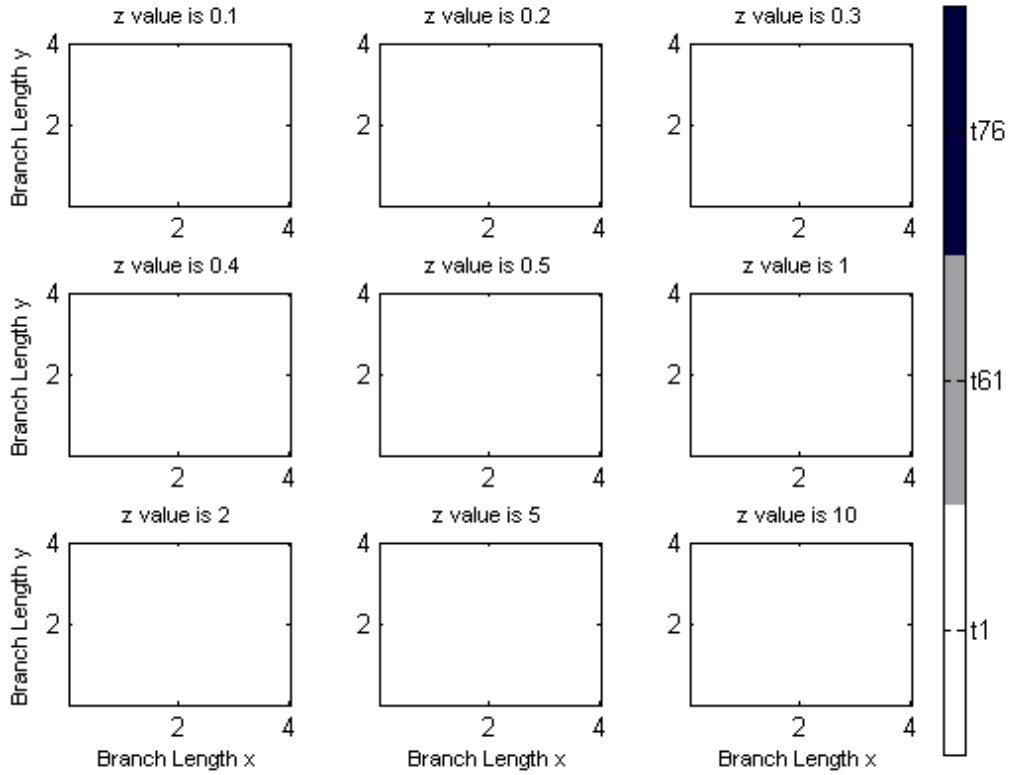


Figure 84. With pruning scheme $0C + 1S$ ($\frac{1}{2}D_1 + \frac{1}{2}D_2$) under the true species tree topology t1. Part (b) is a zoom in of part (a).

(a)



(b)

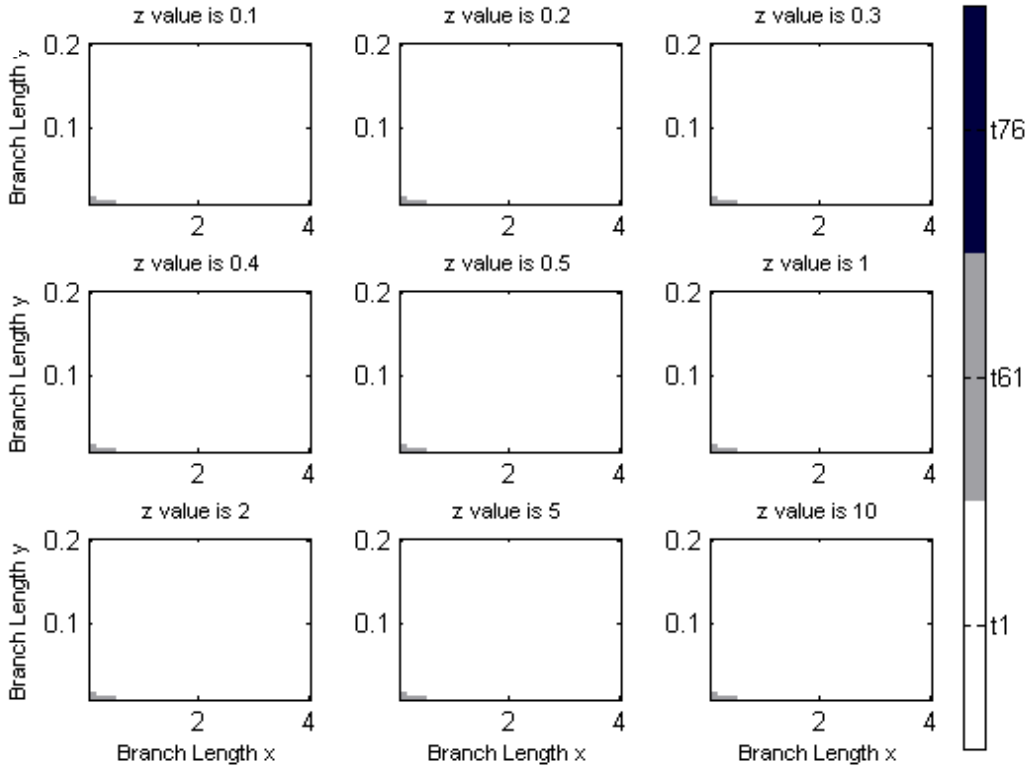


Figure 85. With pruning scheme $0C + 1S$ ($\frac{1}{2}D_1 + \frac{1}{2}D_2$) under the true species tree topology t1. Part (b) is a zoom in of part (a).

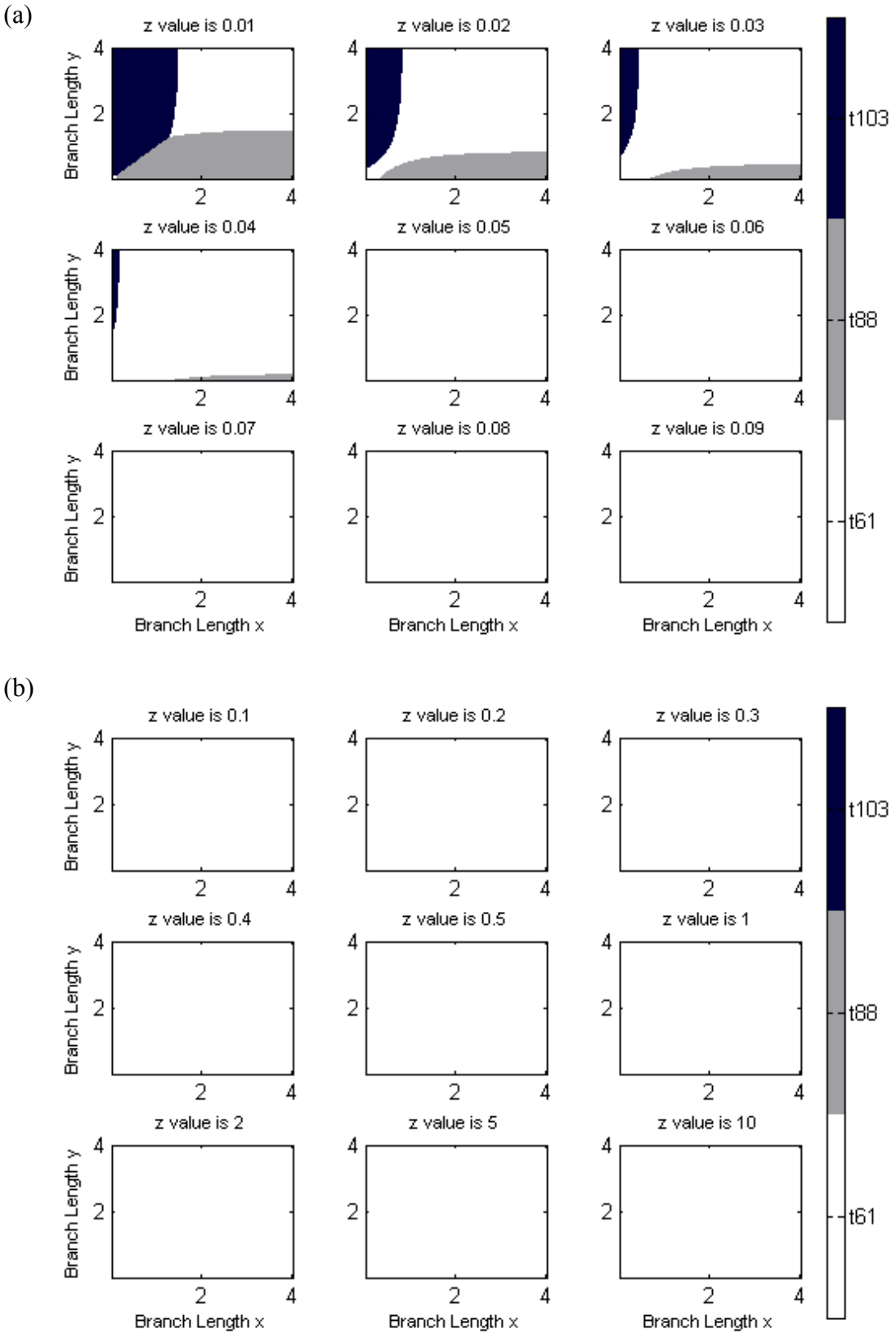
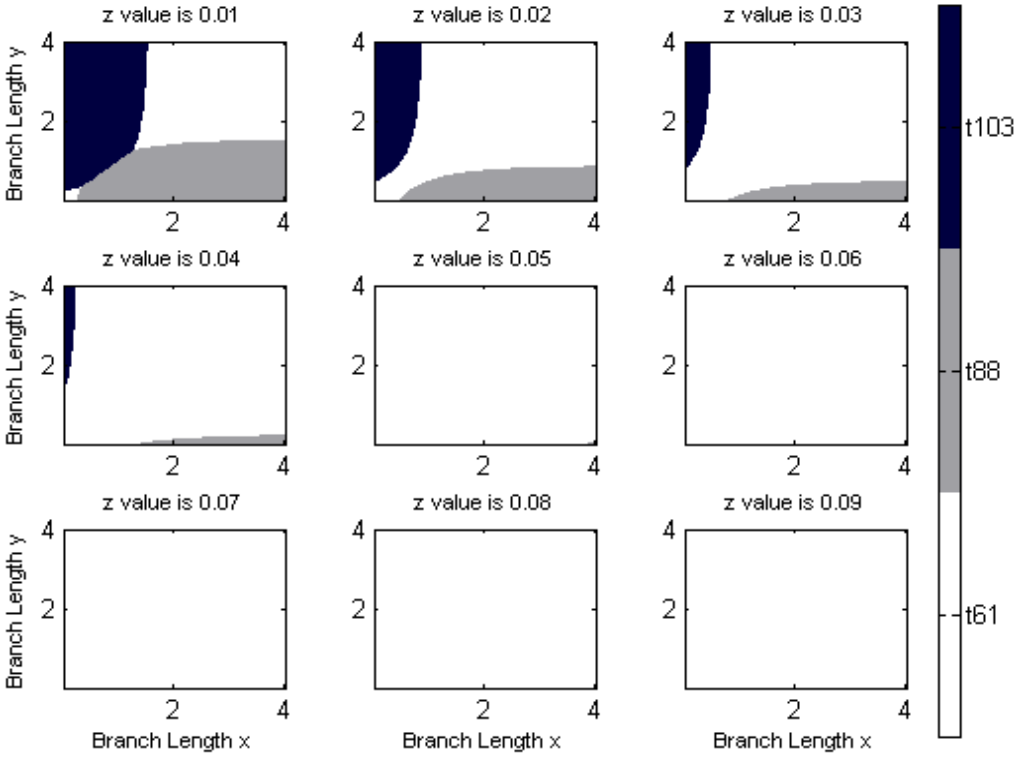


Figure 86. With pruning ½C + ½S (1D_1+ 0D_2) under the true species topology t61.

(a)



(b)

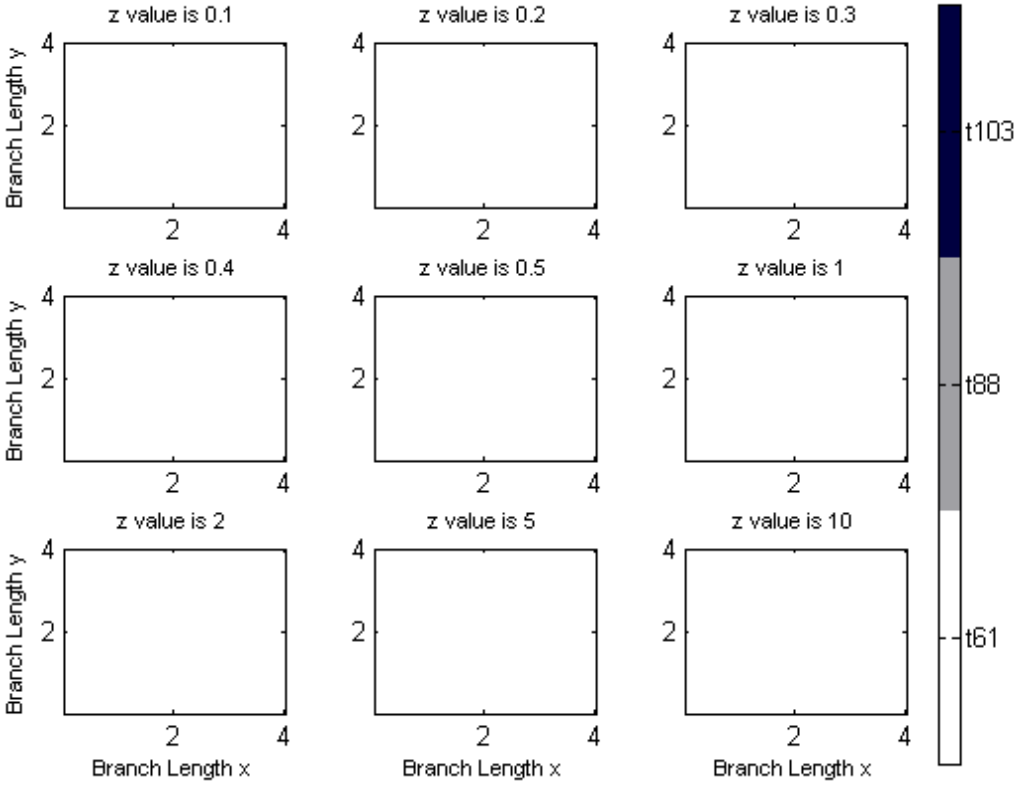


Figure 87. With pruning scheme ½*C* + ½*S* (½D_1+ ½D_2) under the true species tree topology t61.

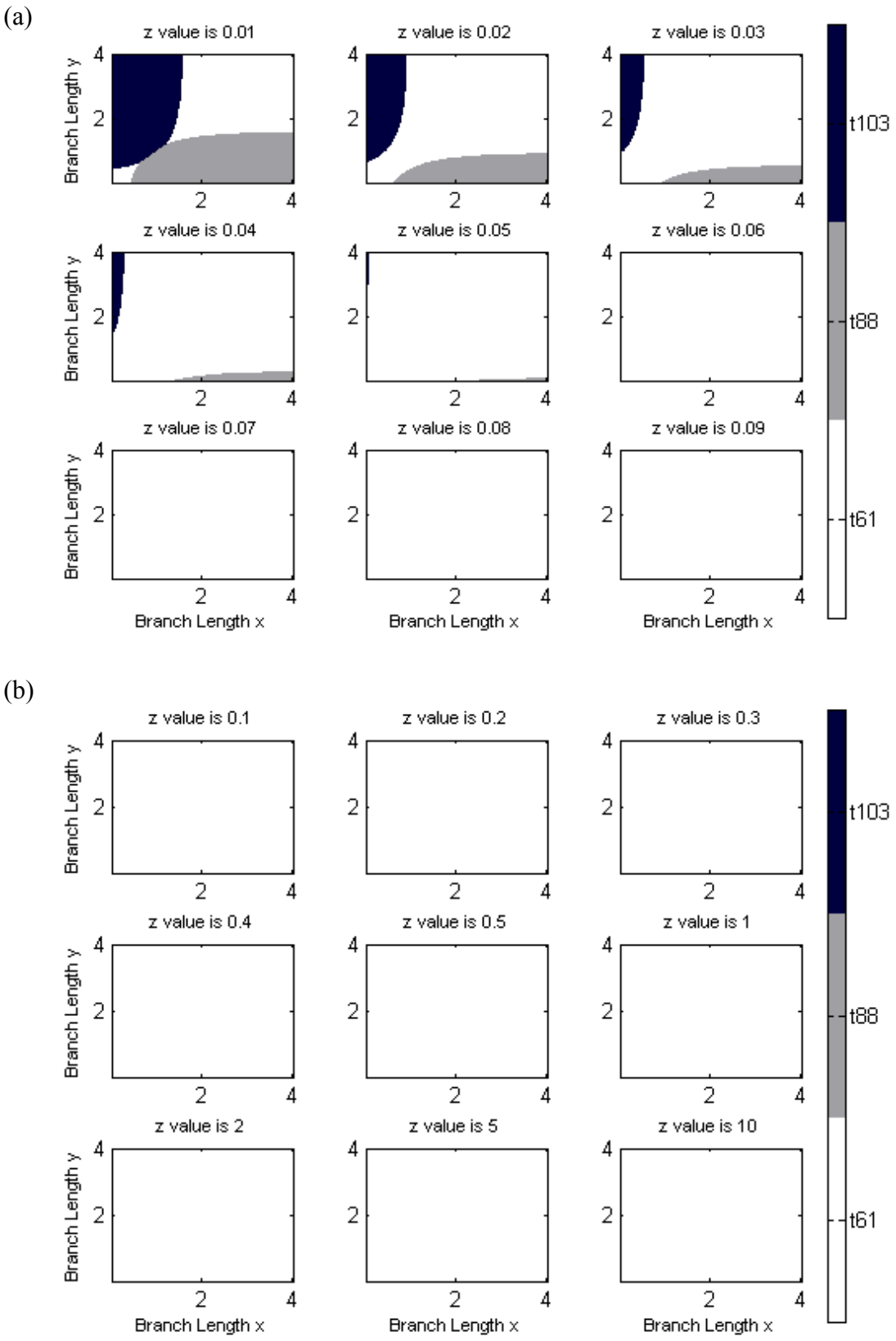


Figure 88. With pruning scheme ½*C* + ½*S* (0D_1+ 1D_2) under the true species tree topology t61.

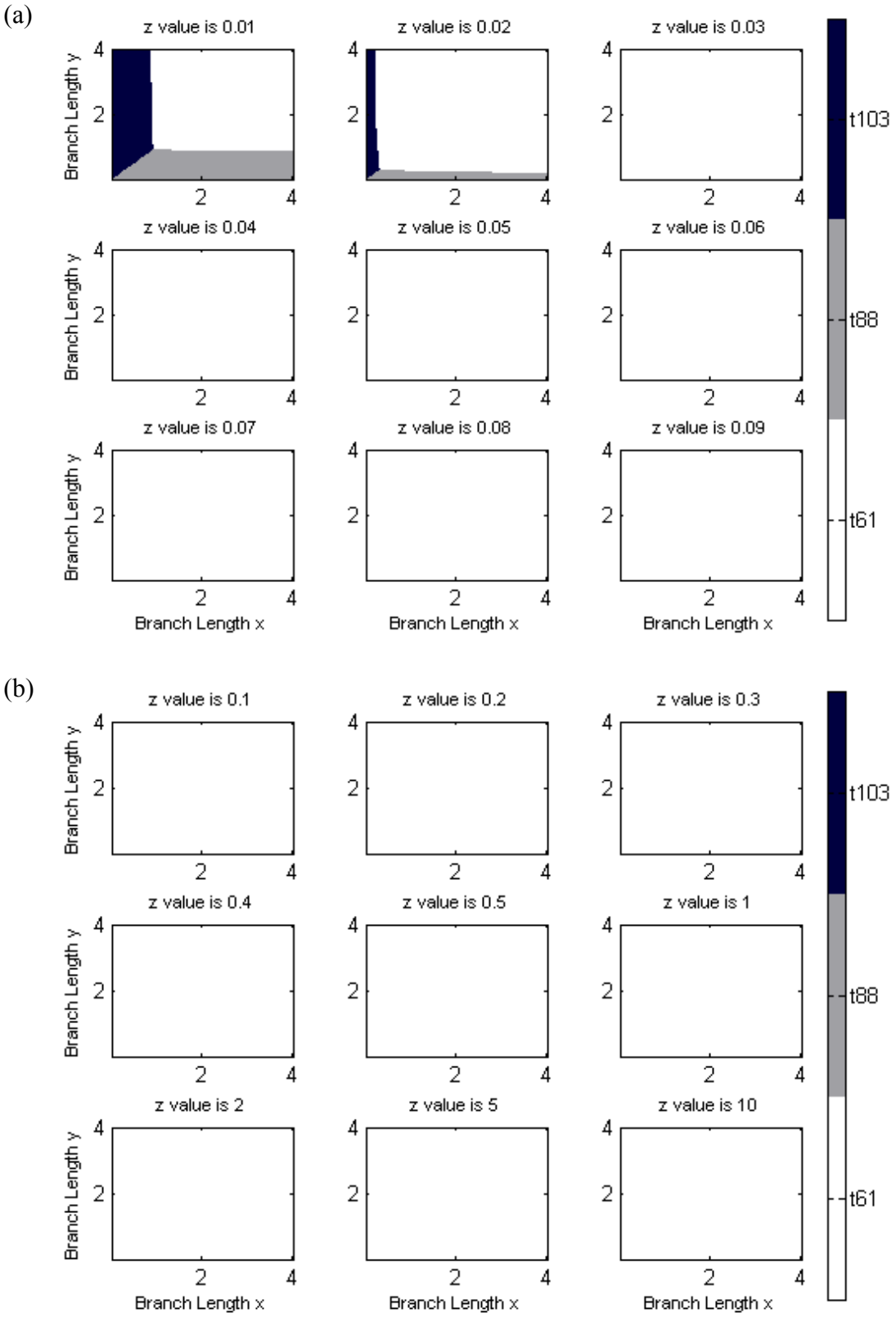
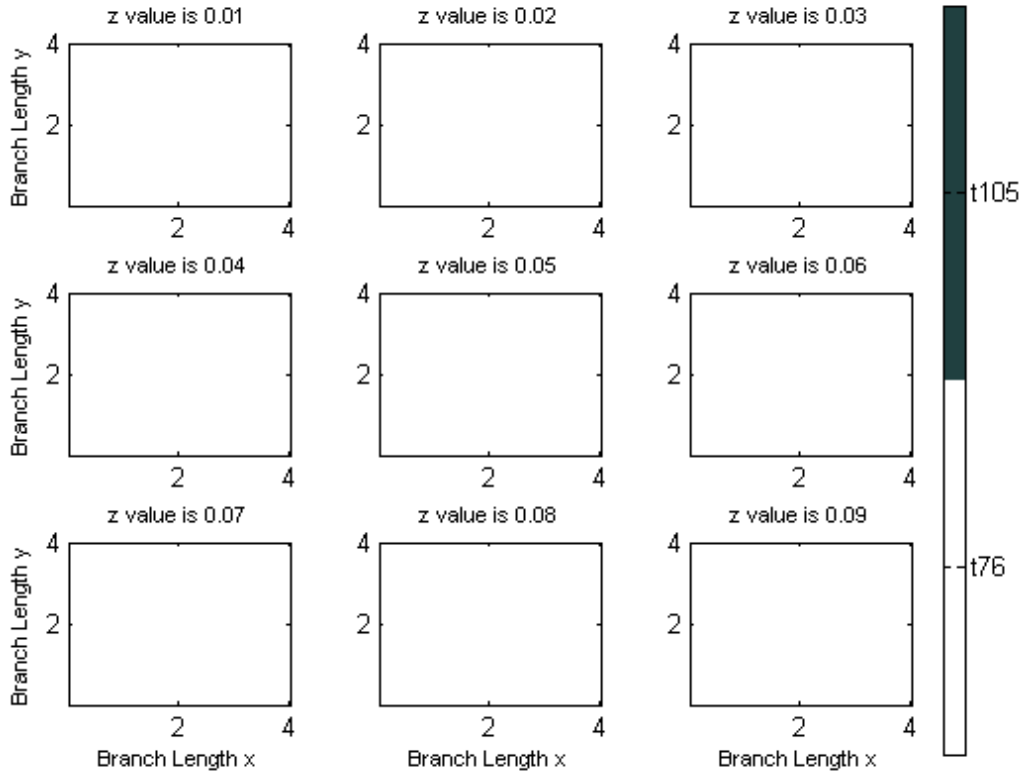


Figure 89. With pruning scheme 0*C* + 1*S* (½D_1+ ½D_2) under the true species tree topology t61.

(a)



(b)

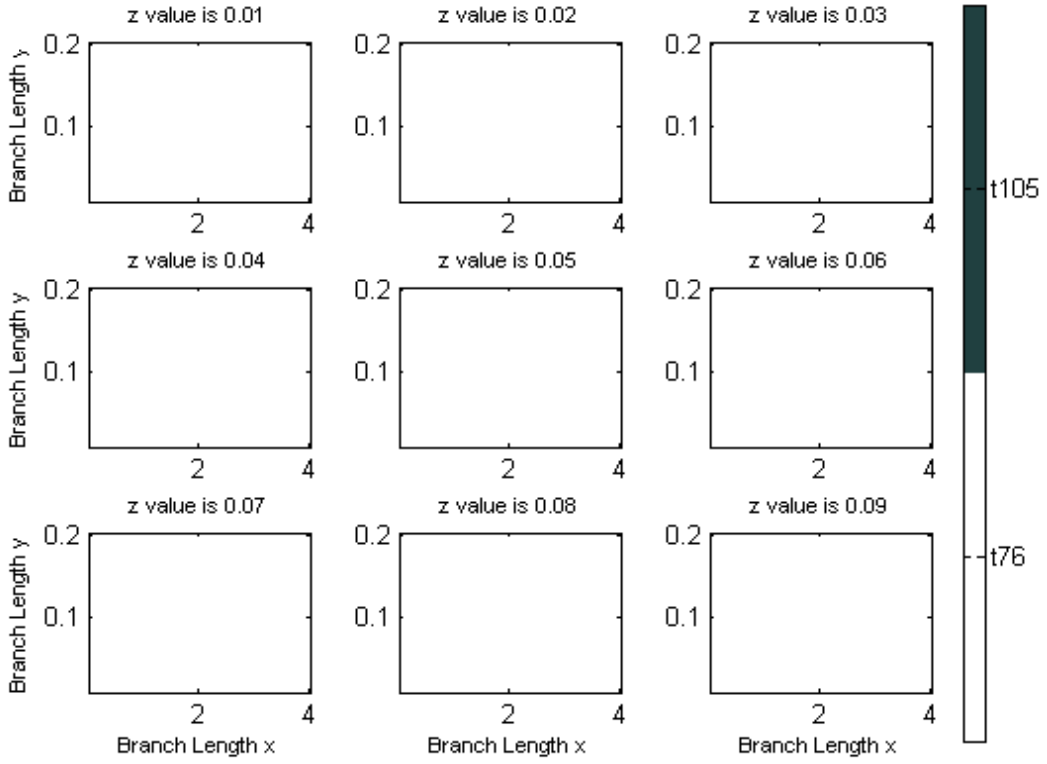


Figure 90. With pruning scheme ½*C* + ½*S* (1D_1+ 0D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

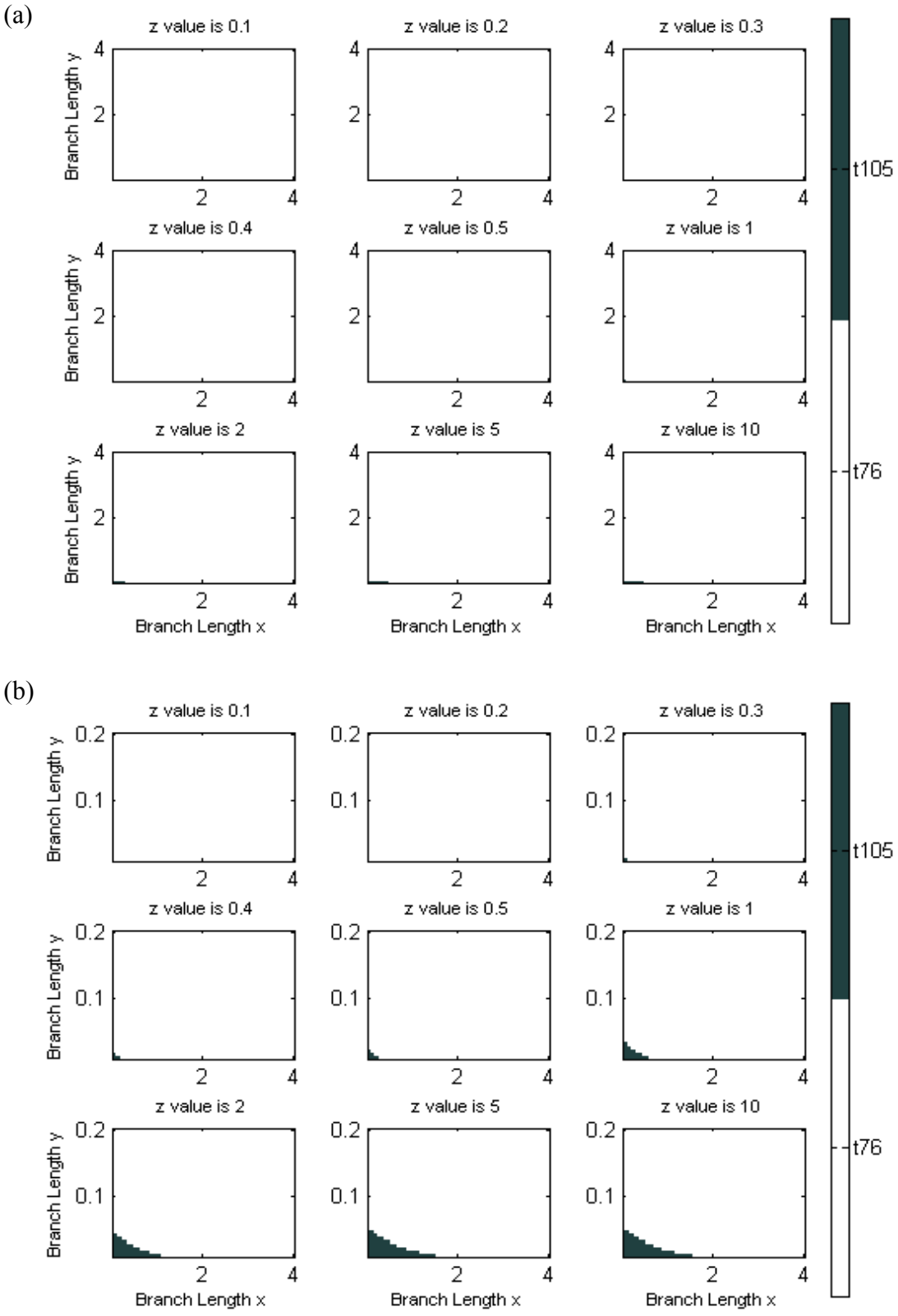
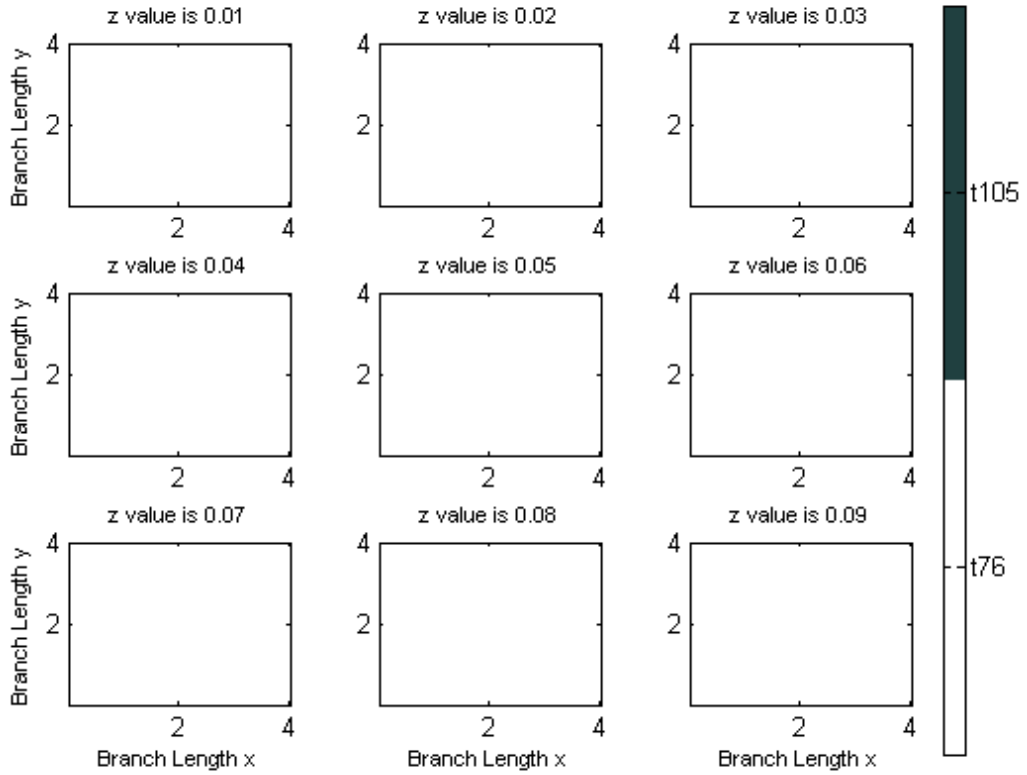


Figure 91. With pruning scheme ½*C* + ½*S* (1D_1+ 0D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

(a)



(b)

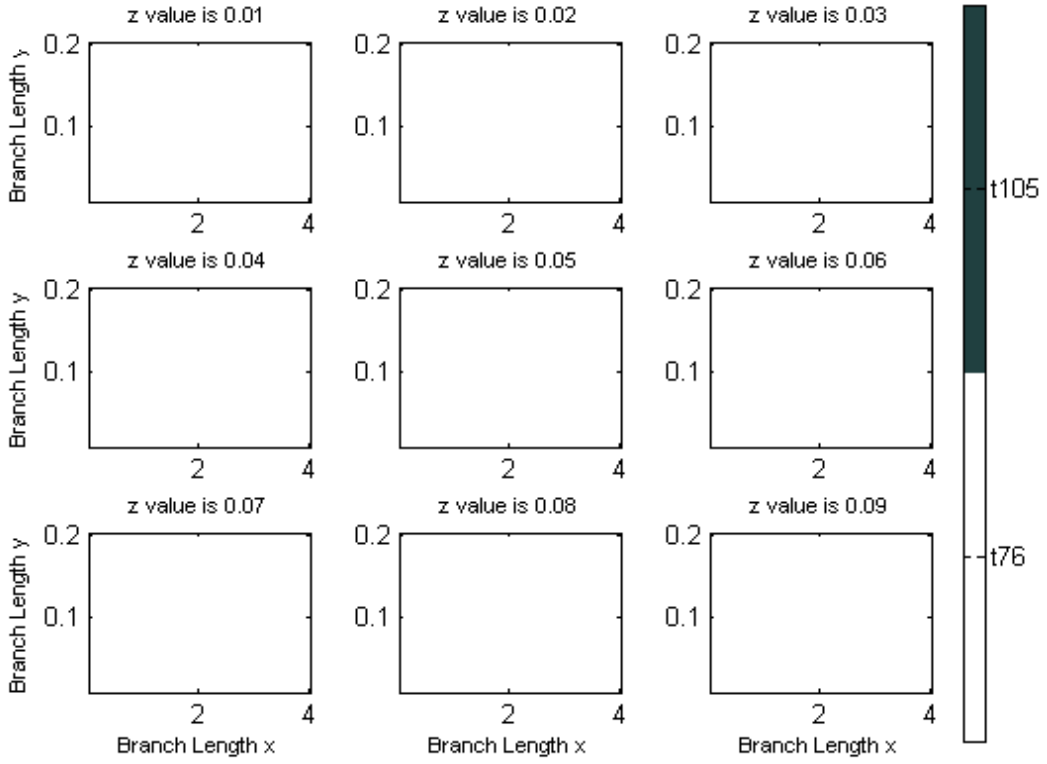


Figure 92. With pruning scheme ½*C* + ½*S* (½D_1+ ½D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

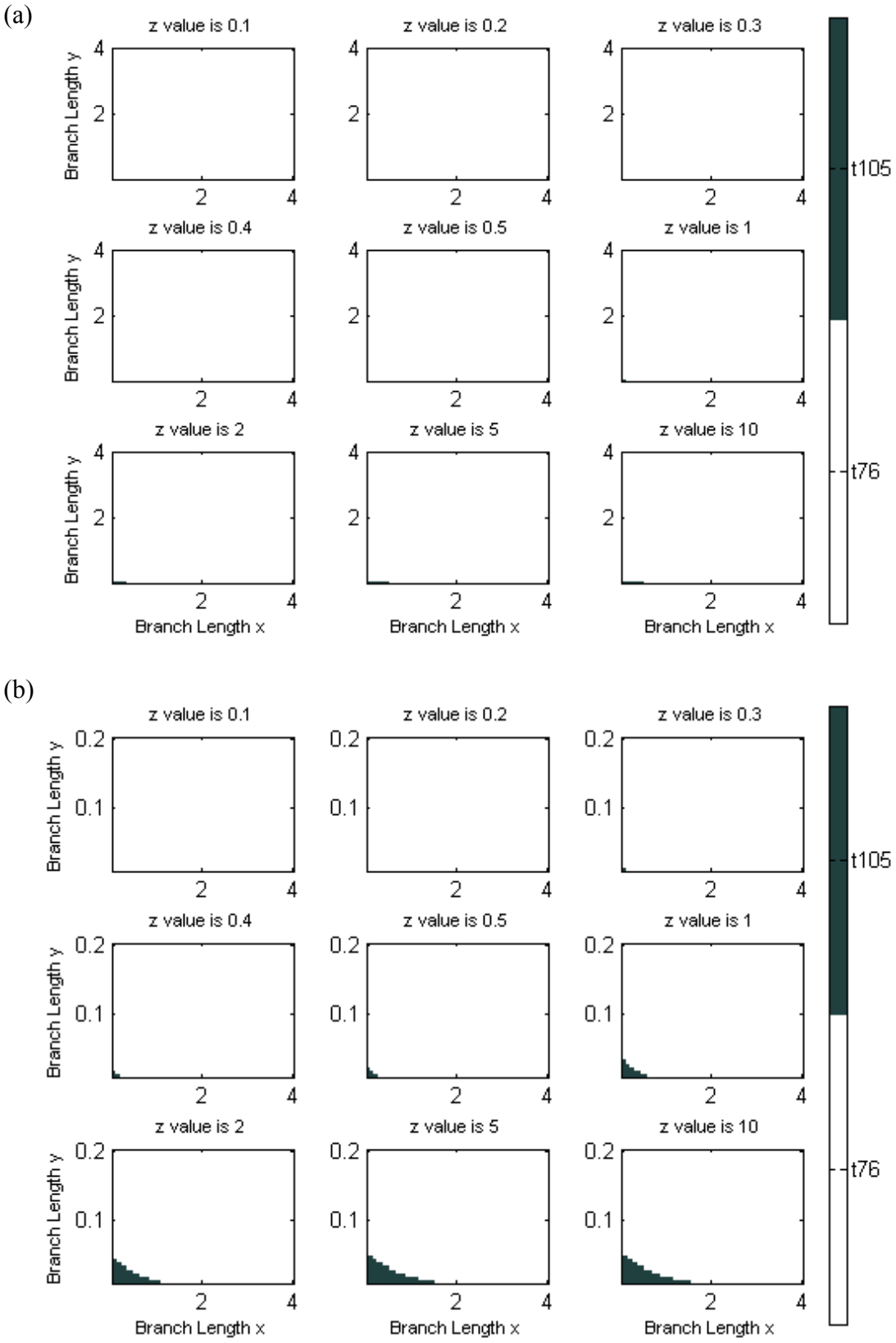


Figure 93. With pruning scheme ½*C* + ½*S* (½D_1+ ½D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

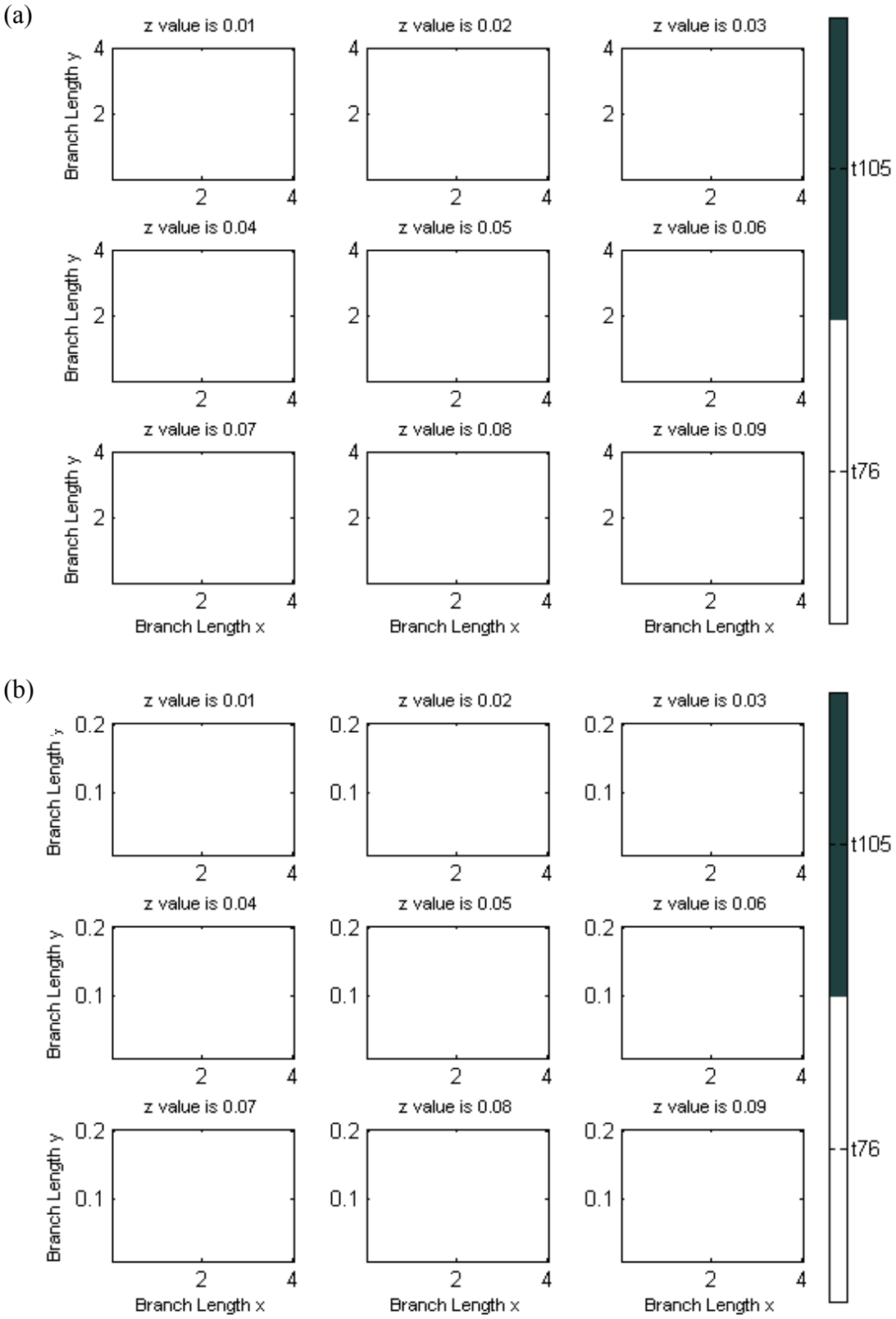
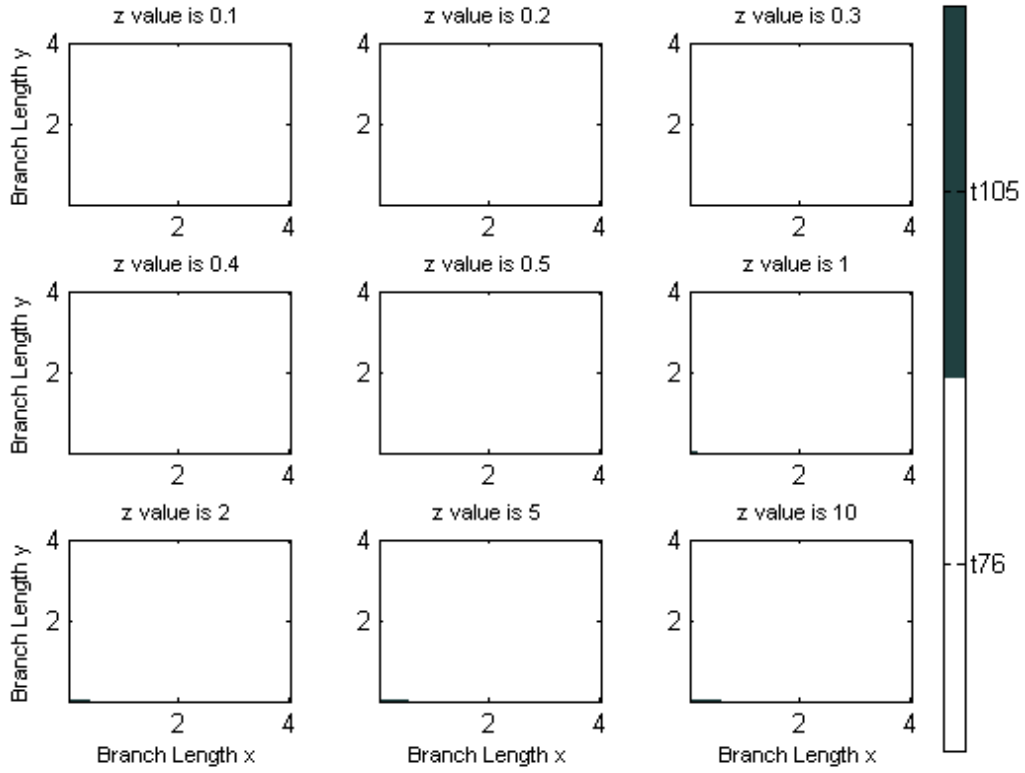


Figure 94. With pruning scheme ½*C* + ½*S* (0D_1+ 1D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

(a)



(b)

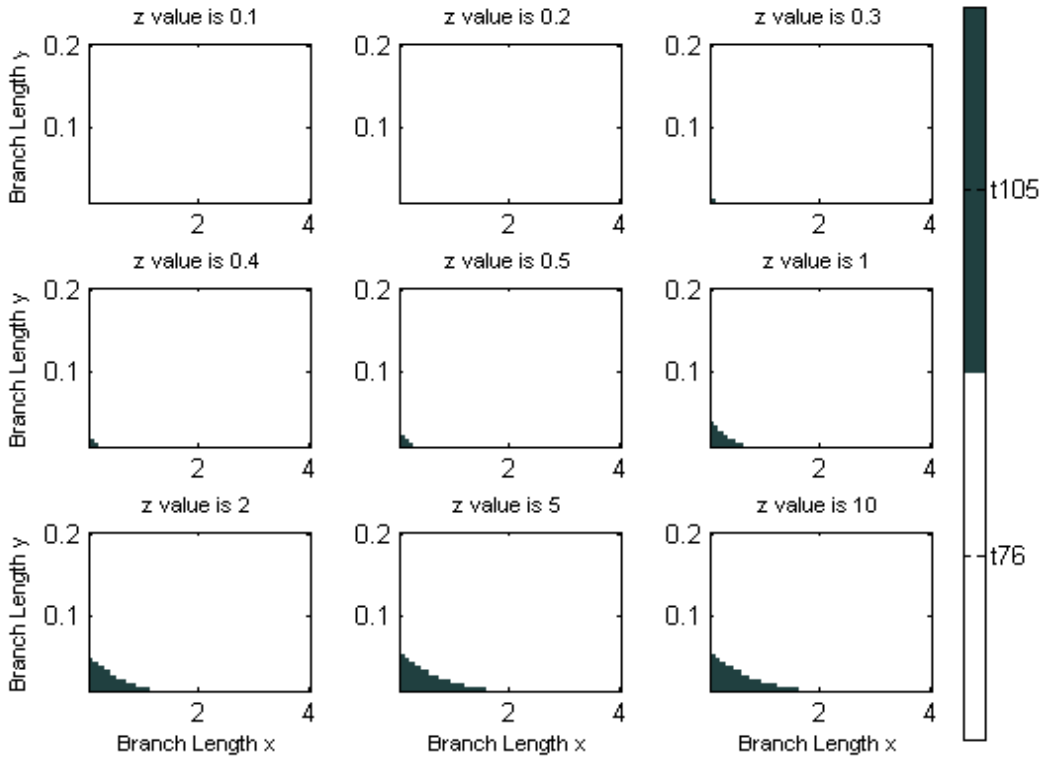


Figure 95. With pruning scheme ½*C* + ½*S* (0D_1+ 1D_2) under the true species tree topology t76. Part (b) is a zoom in of part (a).

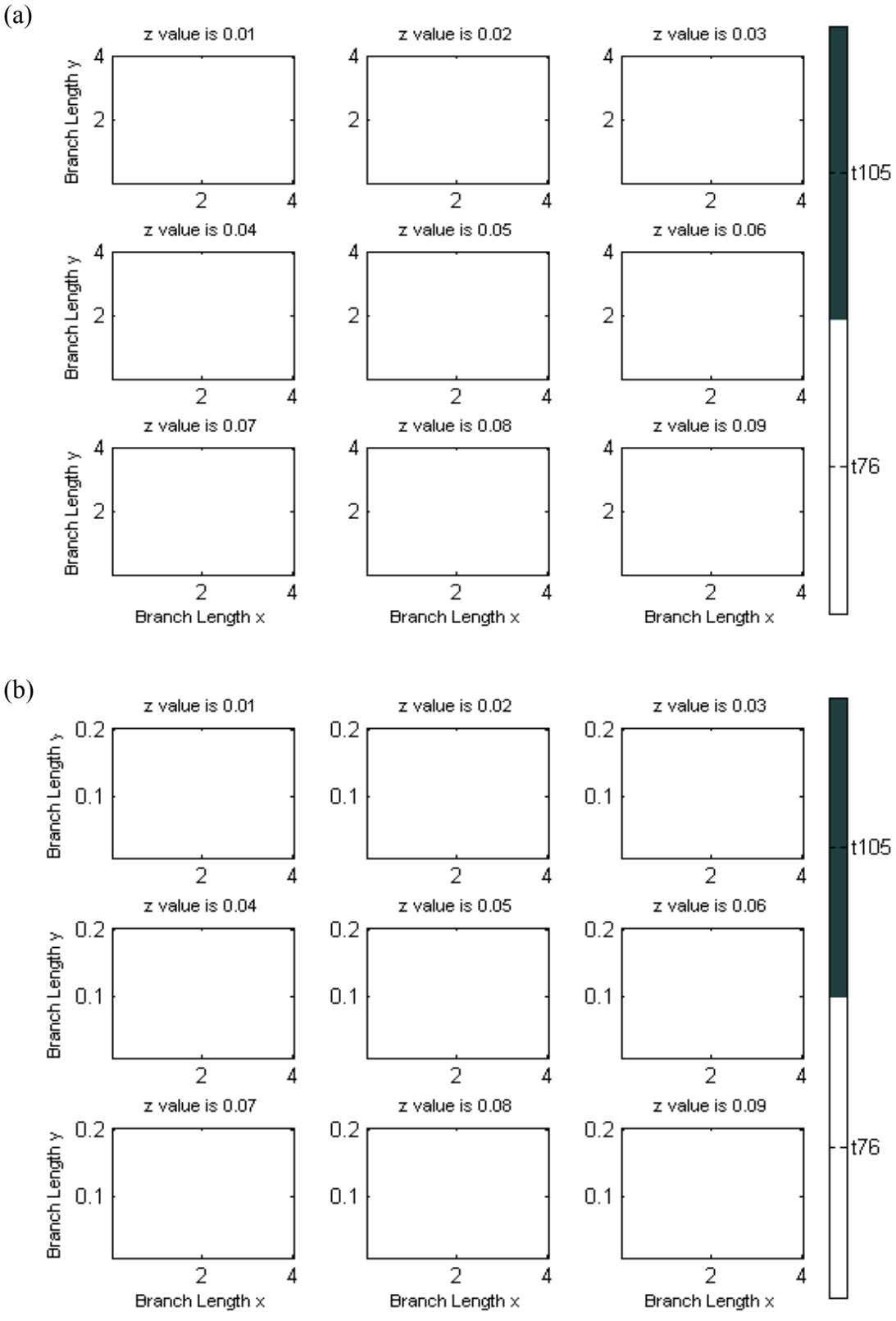


Figure 96. With pruning scheme $0C + 1S$ ($\frac{1}{2}D_1 + \frac{1}{2}D_2$) under the true species tree topology t76. Part (b) is a zoom in of part (a).

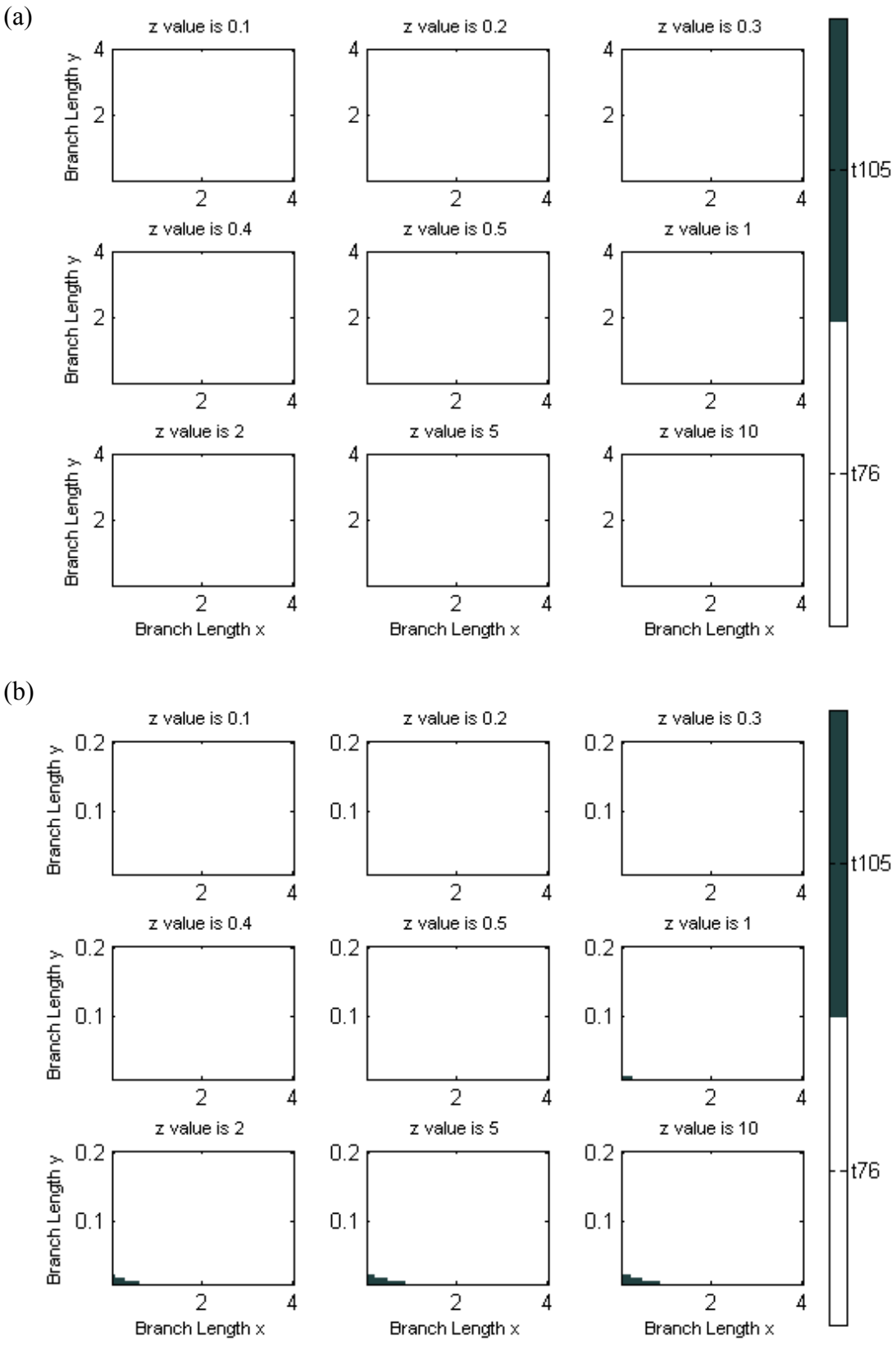


Figure 97. With pruning scheme $0C + 1S$ ($\frac{1}{2}D_1 + \frac{1}{2}D_2$) under the true species tree topology t76. Part (b) is a zoom in of part (a).

References

Baum, B.R. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees, *Taxon*, Vol. 41, 3 – 10.

Bininda-Emonds, O.R.P. (2004). The evolution of supertrees. *Trend in Ecology and Evolution*, Vol. 19, 315 – 322.

Bininda-Emonds O.R.P. and Sanderson M. (2001). Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Syst. Biol.* Vol. 50, 565 – 579.

Bryant, D. 2003. A classification of consensus methods for phylogenies. in Janowitz, M., Lapointe, F.-J., McMorris, F.R., Mirkin, B., Roberts, F.S. (eds) *BioConsensus*, DIMACS. AMS. 163 – 184.

Creevey, C.J. and McInerney, J.O. (2005). Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics*, Vol. 21(3), 390 – 392.

Degnan, J.H. and Rosenberg, N.A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trend in Ecology and Evolution*, Vol. 24, 332 – 340.

Degnan, J.H. and Salter, L.A. (2005). Gene tree distributions under the coalescent process. *Evolution*, Vol. 59(1): 24 – 37.

Degnan, J.H., Degiorgio, M., Bryant, D. and Rosenberg, N.A. (2009). Properties of consensus methods for inferring species trees from gene trees. *Syst. Biol.* 58(1):35-54.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc.

Gatesy, J. and Springer, M.E. (2004). Acritique of matrix representation with parsimony supertree. In *Phylogenetics Supertrees: Combining Information to Reveal the Tree of Life*, Vol. 4, (Bininda-Emonds, O.R.P. ed.), pp. 369 – 388, Kluwer Academic.

Gernhard, T., Hartmann, K. and Steel, M. Stochastic properties of generalised Yule models, with biodiversity applications. In *Journal of Mathematical Biology*, pp. 713 – 735, Springer Berlin, Heidelberg.

Jukes, T.H. and Cantor C.R. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism*. Vol. 3, (Munro, M.N. ed.), pp. 21 – 132, Academic Press, New York.

Klaus, S. (2010). PHANGORN: Phylogenetic analysis in R language. URL http://cran.r-project.org/web/packages/phangorn/index.html.

Maddison, W.P. and Maddison, D.R. (2009). Mesquite: a modular system for evolutionary analysis. Version 2.72 http://mesquiteproject.org.

MAPLE 13. Maplesoft, Inc. Waterloo, Ontario. Canada.

MATLAB (R2008b). Mathwork, Inc. Natick, MA. USA.

Nei, M. (1987). Molecular evolutionary genetics. *Columbia University Press*, New York.

Pamilo, P. and Nei, M. (1988). Relationship between gene trees and species trees. *Mol. Biol. Evol.* Vol. 5, 568 – 583.

Paradis, E., Claude, J. and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*. Vol. 20, 289 – 290.

Ragan, M.A. (1992). Phylogenetic inference based in matrix representation of trees. *Molecular Phylogenetics and Evolution*, Vol. 1, 53 – 58.

R Development Core Team (2009). R: A language and environment for statistical computing. *R Foundation for statistical computing*, Vienna, Austria, (ISBN) 3-900051-07-0, URL http://www.R-project.org.

Robinson D.F. and Foulds L.R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, Vol. 53(1-2), 131 – 147.

Rambaut, A. and Grassly N.C. (1997). Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, Vol. 13, 235 – 238.

Steel, M. and Rodrigo, A. (2008). Maximum likelihood supertrees. *Syst. Biol.*, Vol. 57(2), 243 – 250.

Swofford, D.L. (2002). PAUP*. Phylogenetic Analysis Using Parsimony (* and other methods). Version 4. *Sinauer Associates*, Sunderland, Massachusetts.

Wilkinson, M., Cotton, J.A., Creevey, C., Eulenstein, O., Harris, S.R., Lapointe, F.J., Levasseur, C., Mcierney, J.O. and Pisani, D (2005). The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Syst. Biol.* Vol. 54(3), 419 – 431.

Yule, G.U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J.C. Wllis, *F.R.S.Philos. Trans. R. Soc. Lond. Ser. B* Vol. **213**, 21–87.